CovEx: An Exploratory Search System for COVID-19 Scientific Literature Independent Study Report - Summer 2020

BEHNAM RAHDARI, University of Pittsburgh, USA

This Report presents my attempt to create an exploratory search system CovEx for a collection of academic papers related to COVID-19. CovEx uses concept extraction, knowledge graphs, and user-controlled recommendation to assist users with various levels of domain expertise in their information needs.

1 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 [22]. In this paper we presenting our attempt to augment the set of search systems focused on COVID-19 research literature [24] with a personalized exploratory search system COVID Explorer (CovEx [1]). 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 [21].

We start the paper with the presentation of CovEx interface and follow with the details on concept extraction, knowledge graph organization, and recommendation that enable the work of this interface.

2 RELATED WORK

The *CovEx* system presented in this paper combines the ideas of exploratory search with an important stream of research on personalization, user control, and transparency, It attempts to help researchers discover their *interest profiles* [9], which, in turn, are used to find relevant publications with matching concepts.

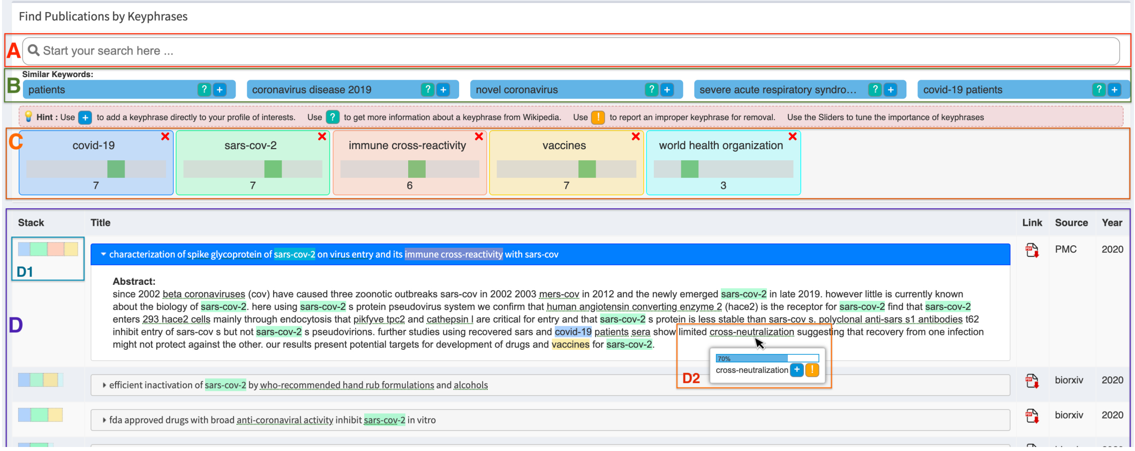
2.1 Exploratory Search

A number of real-life search tasks require a considerable amount of learning during the search process to achieve adequate results. These tasks are known as *exploratory search* tasks [15]. Since simple search systems are usually not efficient in supporting exploratory search, a range of advanced exploratory systems have been developed and evaluated [12, 23]. More recently, few projects in this area demonstrated that the effectiveness of exploratory search could be improved by using a personalized system, which builds a profile of user interests and adapts to the individual user [4, 10]. The work presented in this paper investigates the ideas of profile-based exploratory search in the context of finding research publications related to Covid-19 pandemic.

2.2 Controllability

User controllability has been recognized as a valuable component of advanced information access interfaces. This research was made popular by a stream of work on user controllable recommender systems [13, 16]. However the value of extended user control has been also demonstrated in the area of exploratory search. For example, NameSieve [1] presented a summary of search results in the form of entity clouds, which a controllable filtering and exploration of results. PeopleExplorer [11] offered users an option to re-sort people search results based on multiple user-related

[1] http://scythian.exp.sis.pitt.edu/covex/



Find Publications by Keyphrases

Start your search here ...

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

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

Stack	Title	Link	Source	Year
	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.

cross-neutralization

Fig. 1. Interface Design of Covex representing different parts of the system.

factors. uRank [7] introduced a controllable interface for refining and reorganizing search results. An extension of this work [6] integrated a controllable *social search* into an exploratory search system.

2.3 Open User Profile

The idea to apply open user profiles (also known as open user models) to better support personalized information access was among the early ideas explored in this field. Open user profiles allow users to examine and possibly change the content of their interest profiles, which are used to personalize their search or browsing process. Since the open user profiles increase interactivity, transparency, and controllability of the information exploration process, their application was a good match to the nature of exploratory search. While first attempts to introduce “bag-of-words" open user profiles had mixed success [2], more recent work focused on semantic level user profiles demonstrated its potential for personalized exploratory search [4, 18, 19].

3 THE INTERFACE OF COVEX

Personalized information exploration in CovEx is centered around user interest profile[17] - a collection of keyphrases (keywords) that express user search interests. Unlike traditional search that requires users to specify all keyphrases in a query, CovEx supports users in the process of gradual discovery and refinement of their interests. It also allows the users to control the importance of each keyphrase in recommending relevant results. CovEx interface consists of the following main sections.

Instant Search Box. The search box (Figure 1A) 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 1B). 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 1B). 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. Each recommended keyword also provides users with a short description of the topic. Clicking on the question mark button next to the add button, opens up a separate window containing the abstract of that keyword's Wikipedia entry. This information is crucial when the user is not familiar with the recommended keyword and needs more knowledge to decide whether the keyword must be added to the interest profile.

Slider Area. The slider area (Figure 1C) displays the current interest profile of the user. CovEx implements a content-based recommendation approach, which generates the list of recommended results (Figure 1D) using the interest profile. To support transparency and controllability of this process, the interest profile is visible and directly editable by the end users. To build the profile the user can add relevant topics as explained above as well as remove less relevant keywords (using the red x) as they discover more relevant topics or explore different interests. Sliders associated with each keyword enable users to control the relative importance of a topic compared to others in their profile, ranging from 1 (least important) to 10 (most important). The use of sliders for fine-tuning of user profile was motivated by keyword tuning approach in uRank design [8], which was confirmed as a user-friendly and efficient in an exploratory search context. The initial value of the sliders is set to five but can be changed at any time. All actions within the profile (adding, removing, or adjusting sliders) immediately affects the search results list.

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 keyphrases 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 keyphrase.The second column of table lists the titles of relevant publications. Clicking on each title expands a window that holds the abstract of the paper. The mentioned keyphrases are highlighted with corresponding colors. The opacity of the colors reflect the relevance of a keyphrase to the paper and the current value of slider for that keyphrase. To further assist the users, CovEx underlines all available keyphrases in the text (both in title and abstract). Hovering over the underlined portion of the text opens a popup window (Figure 1:D2) that enable user to (1) see the relevance of the keyphrase to the text in a form of a vertical bar-chart, (2) add the keyphrase directly to the interest profile, and (3) report the improper keyphrases to the administrator for removal. The latter helps us to improve the quality of extracted keywords and eliminate the occasional errors in the process of extraction. Finally, last three columns provide a link to the content of the paper, source and year of publication.

4 THE KNOWLEDGE GRAPH

The knowledge graph consists of three main entities - publications, authors, keyphrases and their relationships - extracted from our data set and hosted in a native graph database Neo4j[2]. 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 keyphrases usign the *Has-Key* relationship. The latter carries a weight that determines the strength of the relationship between each keyphrase and the publication.

[2] https://en.wikipedia.org/wiki/Neo4j

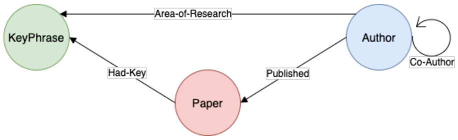


Fig. 2. Graph Schema representing the entities of the knowledge graph and the relationship between them

4.1 Data Source and Graph Statistics

We used COVID-19 Open Research Dataset Challenge (CORD-19)[3] as the main source of data to build the knowledge graph and extract the keyphrases. The dataset contains 51078 document, out of which 48251 documents contain either title or abstract.

Using this dataset and the concept extraction explained below, we generated the knowledge graph covering 48251 publications related to COVID-19 research that have been authored by 157589 researchers. 211862 keyphrases were extracted from titles and abstracts of these publications. Table 1 shows the basic statistics of our knowledge graph.

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

Table 1. Graph Statistics

4.2 Keyphrase Extraction and Weighting

We approach the keyphrase 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 [3]. The standard Bi-LSTM-CRF model consists of three main components, the Embedding layer, the Bi-LSTM layer and the CRF layer. Our implementation of the model is based on the version presented in [14][4]. We obtain the character embeddings of 30 dimensions by training additional Bi-LSTM networks along with the main model. We use the Glove pre-trained word embeddings of 100-dimensions[5]. A 300-dimension hidden layer of LSTM units is used for both the character-level embedding model and the main model. The models are trained using mini-batch stochastic gradient descent with momentum. The batch size, learning rate and decay ratio are set to 10, 0.015 and 0.05, respectively. We also apply dropout to avoid over-fitting and gradient clipping of 5.0 to increase the model's stability.

We train the model with the GENIA dataset[6]: includes 2000 titles and abstracts of scientific articles from Medline database. GENIA is an fully annotated dataset, in which the annotated technical terms cover the identification of physical biological entities (e.g., proteins, cell types) as well as other important terms. We randomly select 300 articles for evaluating and our model achieves 82% of F1-score.

To assign weight for each keyphrase extracted from the the document we found the distance of the keyphrase from the document in embedding space [5]. For training the embedding for concepts extracted from CORD documents we utilized

[3] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[4] github.com/LiyuanLucasLiu/LM-LSTM-CRF
[5] https://nlp.stanford.edu/projects/glove/
[6] http://www.geniaproject.org/genia-corpus/term-corpus

keyphrase embedding [20] and trained the embedding with context extracted from CORD dataset. EmbedRank [5] is used to assign weight to each keyphrase based on the cosine similarity between keyphrase embedding and document embedding.

5 PROFILE-BASED SEARCH

We deploy a two-phase search process to produce the most relevant results based on user interest profile. In the first phase a primary list of candidate have been selected from the graph and the second phase assure that the results are presented to the user in the right order based on their relevancy to the query. We describe these to phases in more details in the following:

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 keywords or tuning the sliders), the system considers all publications connected to at least one of the topics of interest in the user profile. If the number of candidates are less that 20, the system uses similar keyphrases to populate the candidate list. The process of finding similar keyphrases is explained below.

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 keyphrases that appear in the text (title and abstract) of each publication, alongside with their relevancy score (weight) is being generated. Then for every keyphrase that exist in the user interest profile, we multiplied it's 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 1).

$$RelevanceScore_{(f,A)} = \sum_{i=0}^{|A|} Sim_{(a_i,f)} * w_i \tag{1}$$

1. Calculation of relevance score for each candidate publication

In equation 1, A is a set of tuples $\{(a_1, w_1), (a_2, w_2), ...(a_n, w_n)\}$ that represent the current state of the user's profile (topics and weights) and f is a given publication in the graph. a_i and w_i correspond for i^{th} keyword and its slider value at the moment. $Sim_{(a_i,f)}$ shows the value of relevance between a given keyword and a candidate publication in our knowledge graph that has been described in section 4.2

Keyphrase Recommendations. To generate recommended keywords for the current set of keywords in the interest profile, the system generates two sets of candidate keywords using the co-occurrence of seed keyphrase with publications and authors (using collaborative filtering. Then, the system combines the number of co-occurred keyphrases in both sets and uses it as a ranking mechanism. The system presents the top five results to the user.

6 EXPERIENCE AND FUTURE WORK

CovEx system has been deployed online and also demonstrated to several target users. The early results indicate that the success of the system to a considerable extent depends on the quality of keyphrase 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 keyphrase extraction has to consider the collection as a whole increasing user chances to discover keyphrases that could lead to other papers. We are interested to collaborate with experts on keyphrase extraction to develop approaches optimized for exploratory search.

REFERENCES

[1] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Radu Florian. 2010. Semantic annotation based exploratory search for information analysts. *Information processing & management* 46, 4 (2010), 383–402.

[2] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. 2007. Open user profiles for adaptive news systems: help or harm?. In *the 16th international conference on World Wide Web, WWW '07*. ACM, 11–20.

[3] Rabah Alzaidy, Cornelia Caragea, and C. Lee Giles. 2019. Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents. In *The World Wide Web Conference* (San Francisco, CA, USA). 2551–2557.

[4] Fedor Bakalov, Birgitta König-Ries, Andreas Nauerz, and Martin Welsch. 2010. IntrospectiveViews: An Interface for Scrutinizing Semantic User Models. In *18th International Conference on User Modeling, Adaptation, and Personalization*. Springer, 219–230.

[5] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 221–229.

[6] Cecilia di Sciascio, Peter Brusilovsky, and Eduardo Veas. 2018. A study on user-controllable social exploratory search. In *23rd International Conference on Intelligent User Interfaces*. ACM, 353–364.

[7] Cecilia di Sciascio, Vedran Sabol, and Eduardo E. Veas. 2016. Rank As You Go: User-Driven Exploration of Search Results. In *the 21st International Conference on Intelligent User Interfaces (IUI'16)*. ACM, 118–129.

[8] Cecilia di Sciascio, Vedran Sabol, and Eduardo E. Veas. 2016. Rank as you go: User-driven exploration of search results. In *21st International Conference on Intelligent User Interfaces*. 118–129.

[9] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. 2007. User Profiles for Personalized Information Access. In *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer-Verlag, Berlin Heidelberg New York, 54–89.

[10] Amanda Gonçalves Dias, Evangelos E. Milios, and Maria Cristina Ferreira de Oliveira. 2019. TRIVIR: A Visualization System to Support Document Retrieval with High Recall. In *ACM Symposium on Document Engineering*. Article 10.

[11] Shuguang Han, Daqing He, Jiepu Jiang, and Zhen Yue. 2013. Supporting exploratory people search: a study of factor transparency and user control. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 449–458.

[12] Tingting Jiang. 2014. Exploratory search: a critical analysis of the theoretical foundations, system features, and research trends. In *Library and Information Sciences: Trends and Research*. Springer, Berlin, Heidelberg, 79–103.

[13] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and Control in Social Recommenders. In *6th ACM Conference on Recommender Systems*. 43–50.

[14] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank F. Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *32nd AAAI Conference on Artificial Intelligence*. 5253–5260.

[15] Garry Marchionini. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[16] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. ACM, 1085–1088. http://dx.doi.org/10.1145/1357054.1357222

[17] Behnam Rahdari, Peter Brusilovsky, and Dmitriy Babichenko. 2020. Personalizing Information Exploration with an Open User Model. In *31st ACM Conference on Hypertext and Social Media (HT '20)*. Association for Computing Machinery, New York, NY, USA, 0. https://doi.org/10.1145/3372923.3404797

[18] Tuukka Ruotsalo, Kumaripaba Athukorala, Dorota Glowacka, Ksenia Konyushkova, Antti Oulasvirta, Samuli Kaipiainen, Samuel Kaski, and Giulio Jacucci. 2013. Supporting exploratory search tasks with interactive user modeling. In *2013 Annual Meeting of American Society for Information Science and Technology*, Vol. 50. Wiley, 1–10.

[19] Tuukka Ruotsalo, Giulio Jacucci, and Samuel Kaski. 2019. Interactive faceted query suggestion for exploratory search: Whole-session effectiveness and interaction engagement. *Journal of the Association for Information Science and Technology* (2019), In Press.

[20] Khushboo Maulikmihir Thaker, Peter Brusilovsky, and Daqing He. 2018. Concept Enhanced Content Representation for Linking Educational Resources. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence*. 413–420.

[21] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. arXiv:2005.04474 [cs.IR]

[22] Ryen W White, Bill Kules, Steven M Drucker, et al. 2006. Supporting exploratory search. *Commun. ACM* 49, 4 (2006), 36–39.

[23] R. W. White and R. A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan and Claypool.

[24] Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv preprint arXiv:2004.05125* (2020).