Using Knowledge Graphs and Cognitive Approaches for Literature Review Analysis: A Framework

Samaa Elnagar(✉) and Kweku-Muata Osei-Bryson

Information Systems, Virginia Commonwealth University, Richmond, USA
{elnagarsa, kmosei}@vcu.com

Abstract. Advancements in research tools and databases have accelerated the scientific research life cycle. However, the chronological gap between published research, research in progress and emerging research topics is shrinking, thus putting pressure on researchers to find novel research ideas. The Literature Review (LR) process is a fundamental process that can identify gaps in the research literature and stimulate new research ideas. While many researchers adopt different methodologies conducting LR, there is no methodology that can comprehensively unveil innovative research ideas. This research aims to develop a search by concepts framework. The framework involves the use of Natural Language Processing (NLP), Knowledge Graphs (KGs), and Question Answering systems (QA) to ease finding relevant concepts related to a certain scientific topic along with associated files and citations that would in return maximize the efficiency of the scientific research. The framework also allows researchers to visualize the connection between different concepts similar to the cognitive imaging of the human mind.

Keywords: Literature review challenges · Concepts repository · Knowledge graphs · Question Answering systems · Cognitive theories · Information systems research · Thematic analysis

1 Introduction

The emergence of scientific databases such as Google Scholar and literature thematic coding software such as NVivo and Atlas has helped to facilitate the *Literature Review* (*LR*) process significantly [1]. These tools have accelerated the research life cycle by significantly decreasing the time gap between research in progress and emerging topics [2]. These tools also put increased pressure on researchers to promptly publish their research ideas [3]. Unfortunately, the *LR* process is time consuming and usually involves many search rounds to find content that is relevant to the research problem [4]. Moreover, no matter how comprehensive a research process is, there will be some missing captured knowledge [5]. However, knowledge gained by other researchers represents valuable knowledge that is not fully accessible using different databases. Moreover, searching by keywords turned out to be insufficient for achieving a comprehensive depiction of a research issue [6].

M. Themistocleous et al. (Eds.): EMCIS 2020, LNBIP 402, pp. 607–620, 2020.
https://doi.org/10.1007/978-3-030-63396-7_41

According to the *Cognitive load theory,* developed by John [7], the learning process could be enhanced by the proper presentation of information that would lead to superior task performance for individual users. Thematic analysis, specially coding, has been one of the valuable qualitative analysis approaches for *LR* analysis [8]. Coding research literature in the form of codes or concepts specially NVivo codes have been found to be an efficient representation of knowledge [9]. Therefore, sharing other researchers *LR* coding helps realize inaccessible knowledge from traditional research methodologies [10]. Ironically, while Information Systems (IS) researchers have been developing design science artifacts to solve various organizational problems, they rarely develop artifacts that help enhance their research process. We strongly argue that it is the role of IS research to build such an artifact, as the problem strongly relates to IS goal to resolve different society established problems in the light of rigorous technological advancement.

This research aims to build a scientific search concepts framework where previously generated *LR* codes or concepts are considered to be potential idea contributors. This framework is considered to be a design science artifact that could help researchers (designers) to develop innovative research ideas and facilitate the *LR* process. The framework saves the thematic coding of previous research projects in a repository where researchers can search on the concept level. The availability of the repository should accelerate the *LR* research process significantly as a researcher could browse all concepts created by other researchers and pull up the related content associated with those concepts. Searching by concepts offers the opportunity to maximize the value gained through the *LR* in terms of quality ideas through connecting different concepts that might not be accessible using traditional *LR* methods. Additionally, the repository connects different concepts graphically using *KG*s that mimic the human depiction of ideas in the working memory [11].

2 Literature Review Process Challenges

The basic motivation of this research is to overcome the researcher challenges in conducting *LR*. These challenges are external that emerge from limitations in Scientific Databases and software. Other challenges are internal, emerging from limitations in the researcher's cognitive abilities as discussed below.

2.1 Accessibility to Scientific Databases

Despite the plethora of scientific databases, some require paid subscription, which could limit access to important articles and slow down the research lifecycle [12]. However, the research won't violate the publishing rights to paid access. Otherwise, the developed framework will point out the research title where a certain code or concept has occurred. So, researchers will pay only for relevant articles to save time and money to download irrelevant articles.

2.2 Bias of Scientific Databases

Many researchers have reported discrepancies in their research results though using the same keywords [13]. Such discrepancies in results might be as a result of a problem in the *search and retrieve algorithms,* or an intentional behavior in the search engines. For example, a recent article showed how Google interferes with its search algorithms to change results [14]. This issue raises many questions about the significance of the retrieved articles to address the research problem. Building a repository that allows for sharing results of previous LRs, could increase the *likelihood* of finding relevant literature on a research topic.

2.3 Inefficient Search and Retrieve Algorithms

Many search engines are still using ordinary search by keywords and search cues where meta-tags and search keywords are used in retrieving relevant answers. Ordinary search algorithms might not retrieve all relevant answers because they lack semantics and inference in finding search answers [15]. In our research we will be using *Question Answering (QA*) systems that where the input search query is in the form of natural language. The *QA* uses advanced *NLP* algorithms so the output is inferred as candidate answers with confidence score [16].

2.4 Human Cognitive Limitations

Despite the valuable cognitive abilities of humans, they are still limited in their ability to recall and save knowledge. Human working memory is able to process a limited knowledge at a time. "*the limited attention span" is* phenomena where the most frequently visited and the easily accessed areas of the memory surface are activated [17]. Therefore, work developed by other researchers can serve as external memory that aids other researchers [18]. Researchers could benefit from each other's knowledge and expertise by sharing a repository of their thoughts in the form of concepts or thematic codes.

2.5 Approachability of New Ideas

Some valuable knowledge could be found between lines of unreachable research. This knowledge is usually captured by careful reading. However, in a large corpus of *LR*, careful reading of all articles might be challenging. Therefore, accessing concepts coded by other researchers will complement the missing careful reading of important concepts.

Ideas are generated by imaging connected concepts or issues in the working memory [19]. Given the limited cognitive abilities of humans and the large number of concepts found in the extant literature, it is hard to connect all concepts in the human mind which makes the cognitive imaging of ideas *probabilistic* and *incomplete* [20]. This research is using *KG*s to visualize the relationships between different concepts retrieved from the repository [21]. Visualizing concepts relations emulate the cognitive imaging of new scientific ideas.

3 Why to Reuse Thematic Codes?

Thematic Analysis (*TA*) tries to find common patterns (or “themes”) within data. *TA* has the potential to exhibit strong interpretive power through exploring explicit and implicit significance in data [8]. One popular *TA* approach is coding by tagging elements of interest with a coding label in the form of thematic pattern or Nvivo code [22].

TA has significant advantages such as theoretical flexibility for analyzing qualitative data, unlimited dataset size, and popularity across many scientific fields. However, *TA* suffers from limited interpretive power if it is not based on a clear theoretical basis. Therefore, in the design of the framework, we give weights to concepts or codes shared by more researchers because this decreases the subjectivity of a concept and establishes relevance to the topic of research.

4 Theoretical Lens

Kernel theories (prescriptive or a descriptive) help the design of the framework but not to constraint the design and the evaluation of the designed framework [23]. Therefore, kernel theories are used on the conceptual level. The role of theory in this research is to justify and guide the design. The selection of the kernel theories stemmed from the external and internal challenges of the *LR* process discussed earlier that can be summarized to: *accessibility issues, cognitive issues and representation issues*. To maximize the *accessibility* to relevant LR, a researcher should consider other researchers work. We found that the *Transactive memory theory* [24] asserts that other individuals could serve as external memory. This theory reinforces our position that other researchers' *LR*s are essential for generating new research ideas.

In attempts to address the *representation issues* of LR, the *Cognitive fit theory* [25] claims that the presentation of information has a great impact on task performance. This theory supports the application of *KG*s to visualize concepts and their relations that mimic the cognitive imaging process of mind [26]. Representing LR codes or concepts in a form that is similar to how the human mind links different issues will definitely enhance the representation of LR knowledge. Lastly, we have to acknowledge the limitations of human cognition. The *Cognitive Load theory* [27] justify our assumption that human cognition is limited to certain capacity that might limit the depiction of a full image of the research problem. According to the *Cognitive Capacity theory* [28], which accentuate the human limited working memory capacity and the need for assistance to gain a comprehensive picture of the problem. These theories justify the need for external support represented in the concepts generated from previous LR. The depiction of how different theories contributed to the construction of our framework is shown in Fig. 1.

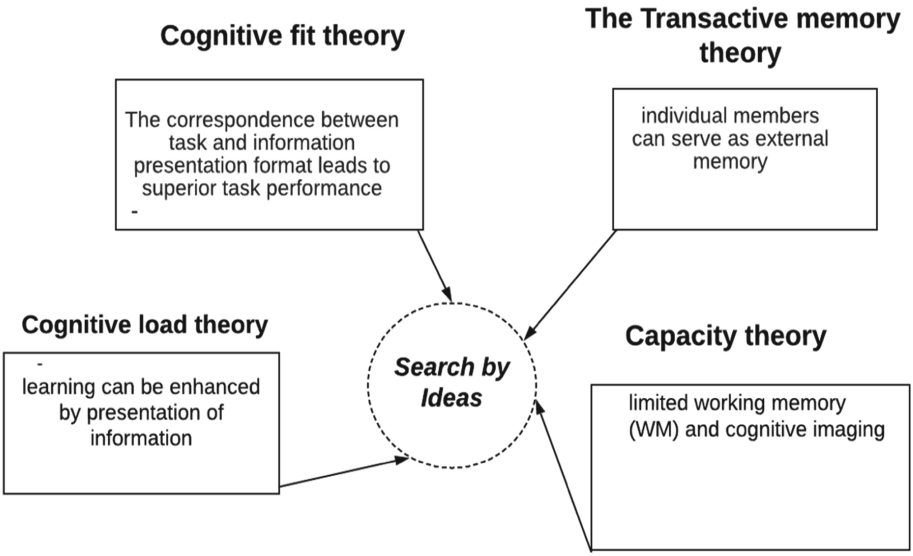

Fig. 1. The Contributing Theories in The Design of The Framework

5 Literature Review

Many previous researches have addressed the challenges in conducting LR, with several previous systems focused on building literature repositories using different types of databases. Those systems adopted the idea of clustering research topics by grouping related concepts using different clustering techniques. To provide an objective literature review and find the gap in literature, we conducted a comparison between previously published research on scientific research search systems as in Table 1.

Table 1. A Comparison between Different Research on Research Search Systems

Research	Objective	Technologies	Domain	Limitations	References
The MIMIC Code Repository, AI for medical	a centralized code base for generating reproducible studies	Relational database, SQL queries, LDA, PCA	Certain Medical domain of critical care, application of AI in medicine	The schema is fixed to certain medical terms, traditional search queries, no import export	[29, 30]
Topic Modeling of Research Fields	Classify/cluster research papers based on topic and language	Topic Modeling, LDA, Naive Bayes, NLP	Domain independent	No repository nor schema, No search and retrieval	[4, 31]
Recommending Scientific Articles	Recommender system for Scientific Articles	Topic Modeling, LDA, CTR,	Domain independent	No schema, limited repository of CiteULike articles, no issues	[6, 32]

(*continued*)

Table 1. (*continued*)

Research	Objective	Technologies	Domain	Limitations	References
Classifying Research Papers with the Computer Science Ontology	Classifier of Computer Science research	Ontologies, SPARQL	Computer Science	Only schema, no systems, no concepts to share	[33]
The Proposed Framework	Search by concepts in previous LR	NLP, KGs, QA systems	Domain independent	Still in initial design phase	

5.1 Gap in Literature

From the previous comparison, we can conclude that no previous related research has targeted the exchange and reuse of *LR* concepts or codes. In addition, most of such research provided schema (ontologies for LR classifying scientific papers) only, third party citation libraries, with no real-time search and retrieval mechanisms. Also, most of such research used *topic modelling* techniques, especially Latent Dirichlet allocation (LDA) to classify or recommend research topics. However, for building a general framework, it is infeasible to depend only on *topic modelling* across different scientific fields. So, the proposed research differs in its objective (i.e. help researchers reaching relevant literature and generate novel research ideas). The research also uses the cutting-edge *KG*s and *QA* systems that involve semantic search and retrieval mechanisms. In addition, the LR concepts should be categorized based on relevancy to the field of study. The framework inspires researchers to generate new research ideas by connecting different concepts stored in a repository of previous LR codes.

6 Research Methodology

We are following a *design thinking* perspective for the design of the *Search by Ideas* framework [34]. This perspective is built upon that a design should go through many design iterations in contact with the actual users of the system. Therefore, the proposed design as an initial iteration where the design is subject to change in each iteration.

To establish the rigor of the research, we built our model upon established IS theories. In addition, we developed a schema that could be saved as an ontology or any structured format such as plain RDF. The summary of the research activities of this research based on [35] design guidelines is summarized in Table 2:

Table 2. DSR Guidelines-based Activities of this Research.

DSR Guideline	Activity of this Research Project
Design as an Artifact	Development of a framework artifact for *Search by Ideas* framework
Design Evaluation	Functional evaluation and performance tests
Research Rigor	Building the artifact based on established theories and the utilization of established research coding schema
Design as a Search Process	Research on *Literature Review* (*LR*) process, *KG*s, *QA* systems and other relevant literature in order to identify appropriate techniques & other results that could be used to inform the design of the procedure

7 The Proposed Framework

The initial structure of the proposed framework is shown in Fig. 2. According to design thinking, the initial design is subject to change in each design iteration and the design should be decomposed into testable components. The framework consists of three main components: the user interface, the controlling system and the repository.

7.1 The User Interface

The user interface is responsible for interacting with users where there are three main functionalities to be supported

1- Interactively ask and answer the user about the research topic and concepts related to research problems.
2- Display and visualize results.
3- Allowing interface for import/export of code projects.

7.2 The Controlling System

We followed modular system design [34] where the system is decomposed to the fine-granular functional units according to design thinking perspective. The units are discussed below in detail.

The QA System: Receives the user questions and convert it into executable query format. The *QA* system is using advanced NLP methods that go beyond topic modelling to hypothesis building and ranking results based on event scoring [36].

The Code Project Parser: the parser search in the imported project for the concepts lists, related files, and citations list if any. In case there is no information provided in the research project description, the parser requests the user to provide the research problem, the scientific field, and the research purpose.

The File Copyright Checker: while parsing the code project file, each file is checked if it has an open access (free access) on the web. If the file has an open access, it will be saved in the repository. Otherwise, DOI and related information will be saved instead.

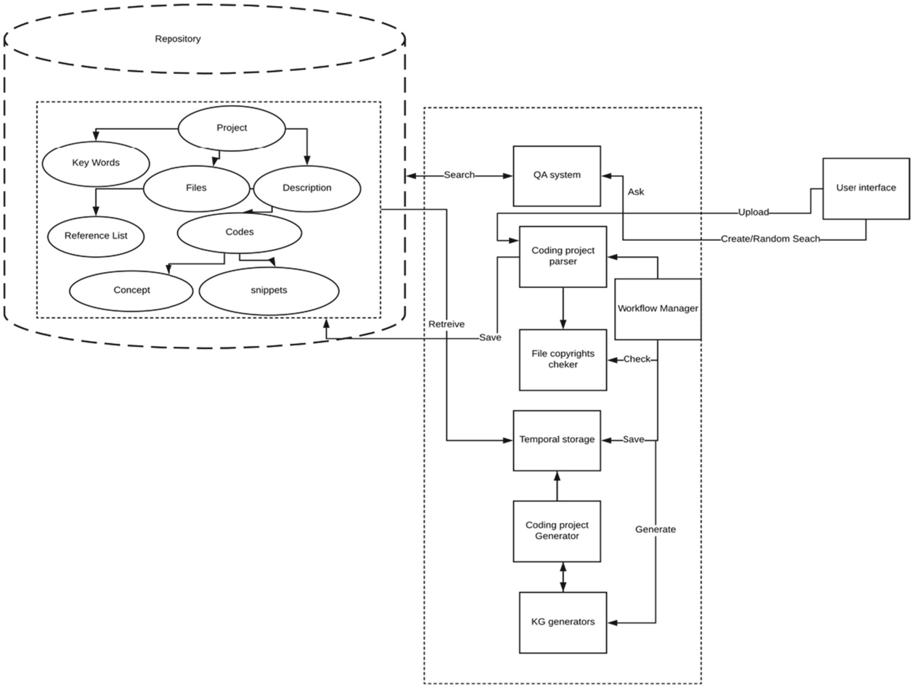

Fig. 2. The search by concepts framework initial design.

The Temporary Storage: it is used to store temporary search results to be visualized or exported as a code project.

The Workflow Manager: is responsible to coordinate between different units in the controlling system and manage the workflow activities *Search, Import, and Export* as shown in Fig. 3.

- *Search activity*: The workflow begins by choosing if the user is conducting a random search or creation of a *LR* project. In both cases, the user asks questions that represent the search queries to be retrieved from the projects repository.
- *The Export activity*: The question answering process will continue till the user is satisfied with the results. Then, the concepts retrieved from the repository along with files and citations are sent to the project generator to be converted to a code project to be downloaded by the user.
- *The Import activity*: In case the user is importing a code project, the project is parsed where the files, concepts, code list and citations are extracted. Then, files are checked for copyrights before they are saved to the coding repository.

The Coding Project Generator is used to convert the search activity codes, files and citations into a code project format so it could be exported and reused.

The KG generator: is used to visualize the relation between the concepts resulting from the search process that are saved in the temporary storage. Visualizing results will help build a cognitive image of the research problem and simulate researcher cognition.

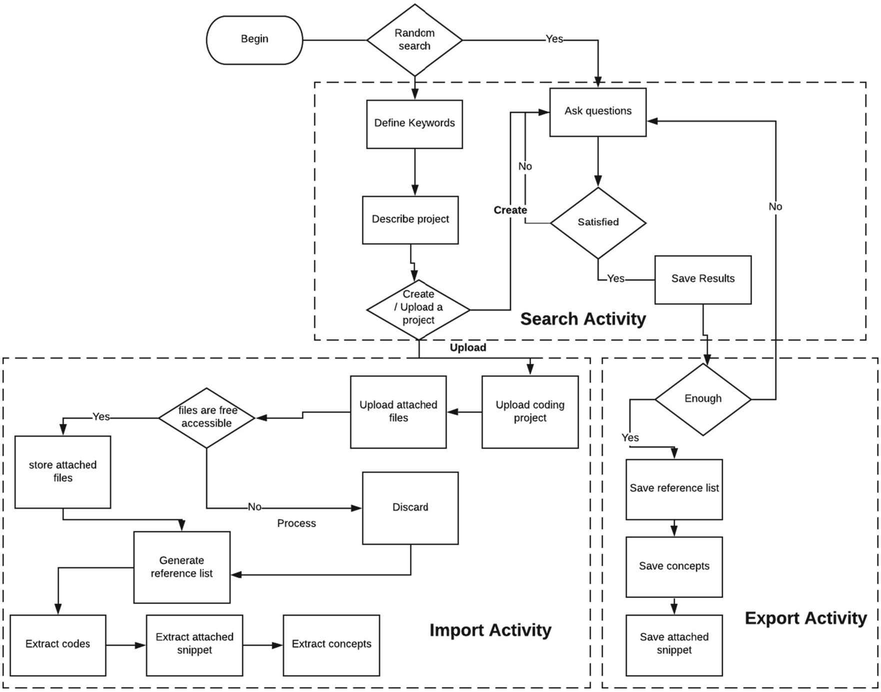

Fig. 3. The Workflow of The Search by Concepts Framework.

7.3 The Code Project Repository

This repository contains the code projects generated by other researchers. However, each code project should be stored in a predefined ontological schema to ease saving and retrieval [37]. The preliminary schema for the code project as shown in Fig. 2 is divided into three classes. The first is the *file class* which contains the list of the files used in the research project. The second class is the *description* class which encompasses the general description of the project and has a subclass of the *reference list* and the *codes* class. The *code class* encompasses the concepts(issues), Concepts weights (how often a concept is coded in research projects), and the corresponding text snippets that represent the concept. The last class is the *keywords class* which represents the meta-data about the research problem of interest to facilitate finding the right context for the search process.

8 Implementation

8.1 Building the Framework Components

The central component in the implementation of the framework is to build the coding repository. In the beginning, we will be using the code projects for a certain software "Nvivo" to avoid compatibility issues. The project itself should be decomposed into files, nodes (codes) with related descriptions that would be saved into the repository. The type of projects will be initially related to information systems research. Then, the design will be modified to include other sciences.

The second step is building the *QA* system. Platforms such as IBM Watson and Microsoft Cognitive services provide *QA* services that could be adjusted according to the project of interest [38]. However, the *QA* has to have full access to the knowledge base or repository to be able to semantically define the context of questions and adjust its performance. In the following step, we are trying to fit the research projects into the coding ontological schema. Then, perform pilot tests to make sure the search and retrieval process is working correctly. Most importantly is representing results from the search process in the form of *KG*s (nodes and edges) to be in the easily readable and simulating for a researcher. Next, *NLP* libraries are used to generate *KG*s of concepts and relations between them. Then, doing a pilot testing for the workflow manager to see if it is well integrated with other components.

8.2 Illustrative Scenario

An IS researcher wants to conduct a literature review about "*Convolutional Networks*" to study how they can be used in real-time object detection. At first, the researcher used the framework interface and wrote his question about *Convolutional Networks* as shown if Fig. 4. The framework allows the user to select multiple scientific disciplines to ensure precise search. The term itself might be used in mathematics and electrical engineering too. Therefore, the framework allows the user to select the main concepts associated with the question. So, The researcher excluded 5G and matrix multiplication from main concepts.

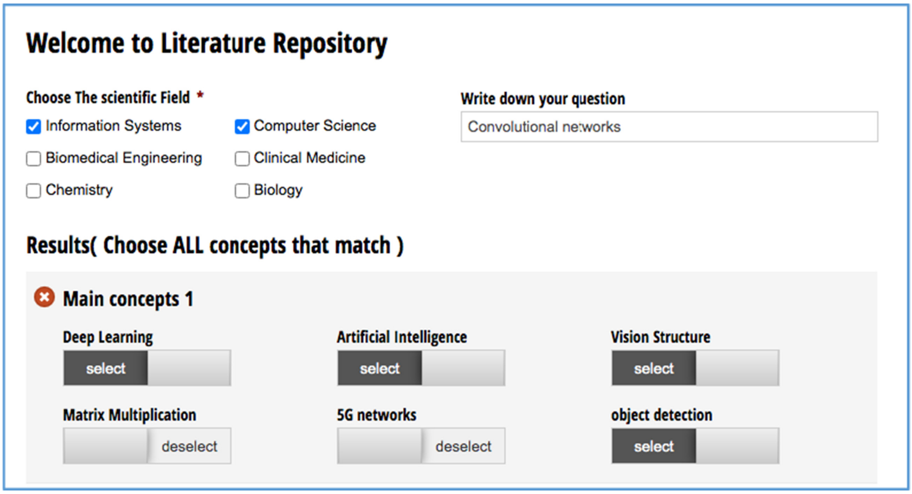

Fig. 4. Search by Concepts User Interface.

Next, the user interface displays available research along with publishing information. The interface also allows quick overview of the places where the concept was coded. If the research file has free access on the web, the researcher can download the file. Next, relevant concepts, parent and child concepts could be included as part of the total search as shown in Fig. 5. Additionally, the researcher could visualize the chosen concepts to be displayed in the form of a *KG* as shown in Fig. 6. A *KG* will display the relationships between concepts along with the weight of each concept (how often a concept appeared in other *LR* coding)

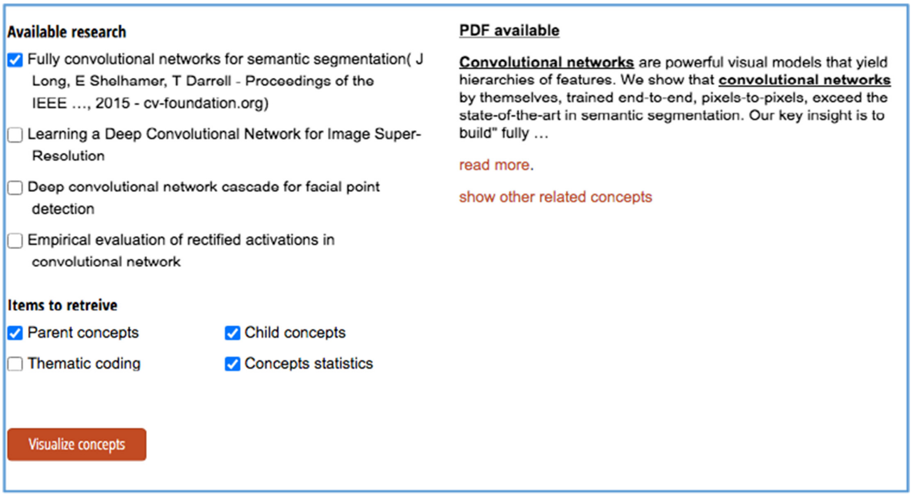

Fig. 5. Search results display.

The *KG* shown in Fig. 6 shows the concepts with strong weight in darker colors than concepts with weak weight or concepts which don't appear often as codes in *LR* thematic coding. We can conclude that the concepts "*Convolutional Networks*" and "*Deep Learning*" are the darkest nodes which appeared mostly in *LR* coding. In addition, the thicker is the line that connects two concepts, the stronger is the relationship between them. For example: "*Convolutional Networks*" and "*Deep Learning*" have strong (thick line) relation while "*Feature Vector*" and "*Activation Functions*" each have a weak (thin line) relation. Lastly, the researcher can decide to export all the search results along with a graph and concepts as a project file.

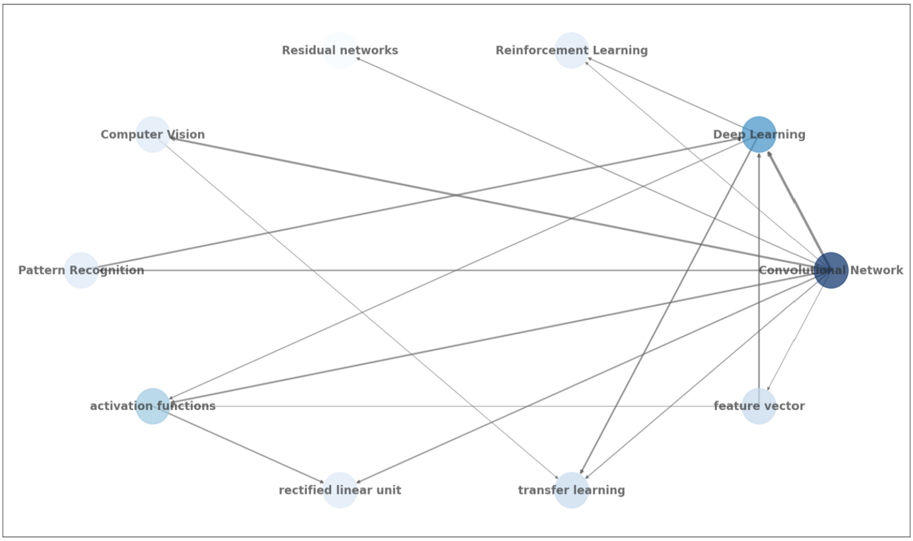

Fig. 6. A Knowledge Graph of Search Results' Concepts.

9 Discussion and Conclusion

This research aims to build a scientific research by concepts framework that reuses the thematic codes and concepts generated from previous research projects. Previous research concepts produced by previous researchers represent the transactive external memory for a researcher that maximizes finding relevant literature. Moreover, previously generated concepts complement the missing concepts needed for the creative cognitive imaging of research ideas. This research aims to build a *Search by Ideas* framework that allows the search, import and export of research projects where the concepts or codes are the central search goal. The retrieved concepts from previous research projects could be visualized in the form of *KG*s that simulates the cognitive imaging of ideas.

The framework is still in the first design iteration. Future work will include the engagement of researchers from different disciplines in the redesign of the framework, so that the design will be modified based on their requirements and suggestions. The evaluation of the framework should be performed by actual framework users. Evaluation should not only be technical/functional but also behavioral based users' feedback on the framework utility.

References

1. Stuckey, H.L.: Methodological issues in social health and diabetes research. J. Soc. Health Diabetes **1**, 56–59 (2013)
2. Luse, A., Mennecke, B.E., Townsend, A.M.: Selecting a research topic: a framework for doctoral students. Int. J. Doctoral Stud. **7**, 143 (2012)

3. Lordkipanidze, R.: Scientific invitation for conducting of researches to protect of honest competition. Int. Charity Sci. Res. Partnership Phys. Persons, **13** (2019)
4. Paul, M., Girju, R.: Topic modeling of research fields: an interdisciplinary perspective. In: Proceedings of the International Conference RANLP-2009 (2009)
5. Brereton, P., et al.: Lessons from applying the systematic literature review process within the software engineering domain. J. Syst. Software **80**(4), 571–583 (2007)
6. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2011)
7. Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. **12**(2), 257–285 (1988)
8. Braun, V., Clarke, V.: Thematic Analysis (2012)
9. Hilal, A.H., Alabri, S.S.: Using NVivo for data analysis in qualitative research. Int. Interdisciplinary J. Educ. **2**(2), 181–186 (2013)
10. Moghaddam, A.: Coding issues in grounded theory. Issues Educ. Res. **16**(1), 52–66 (2006)
11. Nijstad, B.A., Stroebe, W.: How the group affects the mind: a cognitive model of idea generation in groups. Personal. Soc. Psychol. Rev. **10**(3), 186–213 (2006)
12. Salager-Meyer, F.: Scientific publishing in developing countries: challenges for the future. J. English Acad. Purposes **7**(2), 121–132 (2008)
13. Beck, J.C., Perron, L.: Discrepancy-bounded depth first search. In: Proceedings of the Second International Workshop on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR 2000) (2000)
14. Grind, K.: How google interferes with its search algorithms and changes your results. wall street journal 2019. https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753?mod=djemalertNEWS<
15. Büttcher, S., Clarke, C.L., Cormack, G.V.: Information retrieval: implementing and evaluating search engines, MIT Press (2016)
16. Elnagar, S., Osei-Bryson, K.-M.A.: A Cognitive Ideation Support Framework using IBM Watson Services (2019)
17. Findley, N.: What do we mean by 'limited attention span'? Phi Delta Kappan **86**(9), 652–653 (2005)
18. Wegner, D.M., et al.: Paradoxical effects of thought suppression. J. Personality Soc. Psychol. **53**(1), 5 (1987)
19. Colom, R., et al.: Working memory and intelligence are highly related constructs, but why? Intelligence **36**(6), 584–606 (2008)
20. Smith, S.M.: The constraining effects of initial ideas. Group creativity: Innovation Through Collaboration, 15–31 (2003)
21. Elnagar, S., Weistroffer, H.R.: Introducing knowledge graphs to decision support systems design. In: Wrycza, S., Maślankowski, J. (eds.) SIGSAND/PLAIS 2019. LNBIP, vol. 359, pp. 3–11. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29608-7_1
22. Braun, V., Clarke, V.: Successful qualitative research: A practical guide for beginners. Sage (2013)
23. Hart, D., Gregor, S.: Information systems foundations: the role of design science. ANU E Press (2010)
24. Lewis, K.: Measuring transactive memory systems in the field: scale development and validation. J. Appl. Psychol. **88**(4), 587 (2003)
25. Vessey, I.: Cognitive fit: a theory-based analysis of the graphs versus tables literature. Decision Sci. **22**(2), 219–240 (1991)

26. Lin, J., Zhao, Y., Huang, W., Liu, C., Pu, H.: Domain knowledge graph-based research progress of knowledge representation. Neural Comput. Appl. **52**, 1–10 (2020). https://doi.org/10.1007/s00521-020-05057-5
27. Paas, F., Renkl, A., Sweller, J.: Cognitive load theory and instructional design: Recent developments. Educ. Psychol. **38**(1), 1–4 (2003)
28. McCutchen, D.: A capacity theory of writing: working memory in composition. Educ. Psychol. Rev. **8**(3), 299–325 (1996)
29. Johnson, A.E., et al.: The MIMIC Code Repository: enabling reproducibility in critical care research. J. Am. Med. Inf. Assoc. **25**(1), 32–39 (2017)
30. Tran, B.X., et al.: Modeling research topics for artificial intelligence applications in medicine: latent dirichlet allocation application study. J. Med. Internet Res. **21**(11), e15511 (2019)
31. Yau, C.-K., Porter, A., Newman, N., Suominen, A.: Clustering scientific documents with topic modeling. Scientometrics **100**(3), 767–786 (2014). https://doi.org/10.1007/s11192-014-1321-8
32. Pan, C., Li, W.: Research paper recommendation with topic analysis. In: 2010 International Conference On Computer Design and Applications, IEEE (2010)
33. Salatino, A., et al.: Classifying research papers with the computer science ontology (2018)
34. Brown, T.: Design thinking. Harvard Bus. Rev. **86**(6), 84 (2008)
35. Hevner, A., et al.: Design science research in information systems. MIS Q. **28**(1), 75–105 (2004)
36. Bouziane, A., et al.: Question answering systems: survey and trends. Procedia Comput. Sci. **73**, 366–375 (2015)
37. Bunnell, L., Osei-Bryson, K.-M., Yoon, V.Y.: RecSys issues ontology: a knowledge classification of issues for recommender systems researchers. Inf. Syst. Front. **9**(3), 1–42 (2019). https://doi.org/10.1007/s10796-019-09935-9
38. Su, D., et al.: Generalizing question answering system with pre-trained language model fine-tuning. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering (2019)