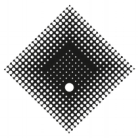

Knowledge graph model development for knowledge discovery in dementia research using cognitive scripting and next-generation graph-based database: a design science research approach

Social Network Analysis and Mining (2022) 12:61
https://doi.org/10.1007/s13278-022-00894-9

ORIGINAL ARTICLE

Knowledge graph model development for knowledge discovery in dementia research using cognitive scripting and next-generation graph-based database: a design science research approach

Kiran Fahd[1] · Yuan Miao[1] · Shah J. Miah[2] · Sitalakshmi Venkatraman[3] · Khandakar Ahmed[1]

Received: 20 October 2021 / Revised: 17 May 2022 / Accepted: 18 May 2022

Abstract
Recent studies report doubling numbers of deaths due to dementia. With such an escalating mortality rate related to cognitive decline diseases, like dementia, timely information on contributing factors and knowledge discovery from evidence-based repositories is warranted. A large amount of scholarly knowledge extracted from research findings on dementia can be understood only using human intelligence for arriving at quality inferences. Due to the unstructured data presented in such a massive dataset of scientific articles available online, gaining insights from the knowledge hidden in the literature is complex and time-consuming. Hence, there is a need for developing a knowledge management model to create, query and maintain a knowledge repository of key elements and their relationships extracted from scholarly articles in a structured manner. In this paper, an innovative knowledge discovery computing model to process key findings from unstructured data from scholarly articles by using the design science research (DSR) methodology is proposed. The solution caters to a novel composition of the cognitive script of crucial knowledge related to dementia and its subsequent transformation from unstructured into a structured format using graph-based next-generation infrastructures. The computing model contains three phases to assist the research community to have a better understanding of the related knowledge in the existing unstructured research articles: (i) article collection and construction of cognitive script, (ii) generation of Cypher statements (a knowledge graph query language) and (iii) creation of graph-based repository and visualization. The performance of the computing model is demonstrated by visualizing the outcome of various search criteria in the form of nodes and their relationships. Our results also demonstrate the effectiveness of visual query and navigation highlighting its usability.

Keywords Design science research · DSR · Knowledge graph · Cognitive script · Dementia · Graph database · Neo4j

1 Introduction

Australian Bureau of Statistics (ABS 2021) report shows that dementia is one of the leading causes of death and the number of deaths has been doubled in recent years. In Australia, the deaths have been increased by 68% over the past decade due to dementia (AIHW 2020). Dementia and Alzheimer diseases are considered the leading cause of death in Australian females followed by coronary heart disease. It is also reported as the second leading cause of death in Australian males next to coronary heart disease. There is no cure for dementia disease; however, the research community is working to develop effective strategies and treatments to reduce the symptoms and improve the quality of life in patients. Such research activities generate a huge amount of knowledge to enhance the understanding of the causes and advancement in the domain but in a document-based form which is understood only using human intelligence. There is a need to gain deeper insights into the scholarly knowledge from research articles (RAs) in this domain to enhance knowledge discovery for a better healthcare system.

Knowledge discovery from the scholarly articles, written in natural language, requires to be automatic and easily understandable with deep insights into the research

✉ Shah J. Miah
shah.miah@newcastle.edu.au

[1] College of Engineering and Science, Victoria University, Ballarat Rd., Footscray, VIC 3011, Australia

[2] Newcastle Business School, University of Newcastle, University Dr., Callaghan, NSW 2308, Australia

[3] Department of Information Technology, Melbourne Polytechnic, 144 High Street, Prahran, VIC 31814, Australia

Published online: 16 June 2022

outcomes such as the modifying risks of dementia in the literature from the biomedical domain. Gaining deep insights into this knowledge is challenging and time-consuming. To encourage and support the exploratory course of action, there is a need to make the research process less tedious from an ever-increasing amount of research. Thus, there is a demand for an automated and effective approach to extract the information from the RAs and to navigate between interrelated information among related articles, which can systematically support the challenging process of discovery. In this context, the recent developments in database management of modern information systems have not been exploited to the full extent, which formed the key motivation of this research. This study addressed the need for a convenient discovery of knowledge from scholarly articles in a structured manner by answering the following research questions:

- Is there a better approach to store or represent the knowledge in the scholarly articles?
- How can a knowledge management model be developed for automatic discovery of dementia risk factors using well-structured next generation infrastructures?

This is a significantly challenging process; however, the proposed approach will be able to minimize the hurdles of discovering knowledge from different sources. Accordingly, this study aims to create a knowledge structure to store extracted core and characteristic aspects of RAs, such as metadata (like title, affiliation) and highlights our research findings. Conversion of the extracted information from the text into a graph-based repository will assist in efficient knowledge discovery and information retrieval as the index free adjacency of native graph storage (NGS) and processing shortens read time (Lal 2015).

We propose an integrated commuting model using next generation graph-based repositories and knowledge mapping techniques to store the extracted knowledge and to achieve semantic visualization of dementia findings from scholarly literature that can contribute to the research community in the domain of biomedical sciences. This computing model adopts a novel approach of combining natural language processing (NLP) techniques, including an innovative cognitive script parser to extract knowledge for creating Cypher and then store it into a graph-based structured repository for efficient storage, query, and management such as Lal (2015). The cognitive script intelligent parser is developed in Python to read the cognitive script and generate Cypher statements. Further, it incorporates graph representation using knowledge maps towards catering to the requirements of the next generation digital era. The model facilitates easy navigation through the literature along with visual querying of non-relational graph-based repositories. Our modelling approach is unique for representing a complex relationship of knowledge from scholarly articles, with a proof-of-concept implementation of an innovative modelling approach applied to dementia research articles. The main objective of the study is to formulate an advanced approach to extract the knowledge entities and their hidden relationships from scholarly articles as well as store the extracted knowledge in a repository for future retrieval and deep understanding of the domain topic. The key contribution of this study is a proof-of-concept framework for developing an automatic knowledge discovery from scholarly articles to create a permanent graph-based repository and to facilitate easy navigation and querying. The implementation of the model combines the data and knowledge mapping techniques for extracting the relevant knowledge from domain-specific scholarly articles. We employ advanced techniques of text mining to identify meaningful entities and their relationships to create a graph-based repository for knowledge discovery and visual queries.

The rest of the paper is organized as follows. First, we provide an overview of related works as a literature review in Section II. The adoption of design science research (DSR) methodology for the development of our proposed computing model is introduced in section III. Section IV illustrates the application of the computing model in dementia using multiple case examples. Section V gives a summary of results with discussion and evaluation. Finally, in Section VI, the paper provides conclusions and current limitations that motivate researchers for future work in this research direction.

1.1 Literature review

In the research community, RAs are the most complete representation of data of human knowledge. There is substantial growth in the publications of RAs (Kertkeidkachorn and Ichise 2018). While technology has significantly advanced, the research process for novel findings has not transformed much. Bridging the gap to create structured knowledge from unstructured or semi-structured text obtained from RAs on dementia will open an utterly new world of possibilities for the research community in this domain. This section provides a review of selected literature that is related to the components of this study.

Multiple researchers agreed (Hansen and Kautz (2004) cited in Balaida et al. (2016), Tiwana (1999), White (2002) that there are advantages of using a knowledge map, such as effective storage and traversal of knowledge represented by the knowledge map. Researchers Kaur and Chopra (2016), Aslam et al. (2017) concluded that unstructured data in the form of text are very complex; therefore, powerful techniques are required to extract the data. Most text mining tools used NLP techniques and in some cases the combination of machine learning techniques. Some examples of such

tools are Open Calais, Rapid miner Text Mining and SAS text miner. The major finding is to extract the information about the RA, like title, author, year of publication, publisher and editors, and record it in a database of a library to create an open library publication framework.

An existing tool Semantic Scholar understands the knowledge given in the RAs (AIHW 2020). It uses AI to help retrieve the most related information quickly and efficiently. It extracts the limited information from RAs, which is metadata of the RA. For example, information extracted includes the title of the RA, abstract, author details, citation, keywords, and references from the RA. However, it does not represent the actual knowledge provided by the research (Semantic Scholar 2021). Another similar commonly used tool is Google Scholar (Google 2021).

Di Iorio (2015) established the standard RAs in simplified HTML (RASH) framework, which is a markup language that uses 31 HTML elements while academic RAs use only 25 elements for writing articles online. The HTML5 elements are: a, blockquote, body, code, em, figcaption, figure, h1, head, html, img, li, link, math, meta, ol, p, pre, q, script, section, span, strong, sub, sup, svg, table, td, th, title, tr, ul. This framework provides a transformation from the semi-structured to the unstructured nature of a RA. Gardner et. al. (2018) describe an NLP library AllenNLP to support programming for NLP research by providing deep learning methods. AllenNLP has reasonable performance for simple text used in general purposes but is not suitable for RAs. Kertkeidkachorn and Ichise (2018) proposed a text to knowledge graph (T2KG) framework, which automatically creates a knowledge map from unstructured text. T2KG creates the knowledge graph by using entity mapping, coreference and triple extraction. Similarly, a semantic graph capturing the evolution of RAs is introduced Bayram et. al. (2021). On the other hand, the FRED tool (Gangemi et al. 2017) creates its own ontology linking existing knowledge entities to other existing knowledge bases, such as DBpedia. The ontology created is metadata oriented. It does not deal with the article content but reports only the categorization of the articles (Table 1).

Programming languages like Java (Manning 2014) and Python (Bird 2009; Qi et al. 2020) provide NLP techniques and tools. Python, a scripting language, offers powerful built-in idioms for NLP. A Python-based NLTK is a collection of libraries, which offers NLP techniques like tokenizing, stemming, tagging and semantic reasoning (Bird 2009; Maksutov et al. 2020). DyNet is an NLP programming toolkit (Neubig et al. 2017), which implements a neural network model with a dynamic declaration strategy. A java based annotation tool Stanford CoreNLP provides NLP techniques like tokenizing, sequence split, part of speech tags a NER (named entity recognition) of entities like PERSON, LOCATION, ORGANIZATION) through to coreference resolution (Manning 2014). Open Calais API is another example of an NLP tool developed in Java. It automatically extracts semantic information from the unstructured text and annotates the text as the people, places, and events. OpenCalais is a web-based tagging engine for text processing for tagging of industries, topics and social tagging in unstructured text. It also provides a facility to generate a resource description framework (RDF) file, which is a structured form of knowledge (Gangemi 2013). Existing works report a few pre-trained language representation models and consider state-of-the-art methods like bidirectional encoder representations from transformers (BERT) (Devlin 2019) or generative pre-trained transformer 3 (GPT3) (Zhu and Luo 2021) that can overcome the limitations of other language models. Some language models such as WEC (word embedding-based clustering) (Comito et al. 2019) to detect topics show significant accuracy among state-of-the-art methods. However, the limitations of these models are: i) restriction of the search information to only text, ii) requirement of a complex setup and ii) lack of a permeant repository creation. These gaps in the literature form the motivation for our unique contribution to this research work.

Markus et al. (2015) cited in Paulheim (2016) confirmed that links between different knowledge maps can fill the missing knowledge gap. In addition, Paulheim (2016) defined a DeFacto system that converts the statements in DBpedia to human language text. Considering the creation of the graph-based repository of knowledge obtained from RAs on dementia, Lam et al. (2007) have generated structured RDF format-based ontology from different sources of neuroscience, which is a framework Alzpharm, which is considered as former generation of NLP work, that integrates the neuroscience information from RDF format as single ontology.

Most of the existing models and tools, as reviewed above, require a complex process of system installation and focus on tagging, annotation, tokenizing and NER-based NLP techniques that output as representations only. They do not facilitate the storage and retrieval of the semantic analysis of the semi-structured text. However, this study used the creation of a cognitive script to extract the semantic meanings of the findings given in natural language text for creating a graph-based repository to facilitate the analysis by using an intelligent Python script. The dataset of the cognitive script from the sample research articles consists of unstructured data, and the techniques employed are different from other related studies that apply data integration techniques using grids and peer-to-peer networks on distributed or heterogeneous datasets (Comito et al. 2007, 2004). We develop an innovative Python program to generate Cypher statements from cognitive script to create nodes and edges in Neo4j. Furthermore, queries are developed by using SQL in existing works (Comito et al. 2007), while our study employs

knowledge map modelling queries, which offers visual output to facilitate easy navigation and deep understanding of the domain topic.

Knowledge mapping in the domain of biomedical sciences is based on creating ontologies or linking the structured knowledge entities, in the form of RDF, to ontologies, whereas this study extracts the entities and their relationships from the text and converts it into knowledge for storing it in the graph-based repository. We intend to match the knowledge within the text from a collection of abstracts of RAs on dementia, which may even describe information that has not been captured yet, due to its novelty of unearthing hidden linked knowledge. While existing works reported in the literature can generate knowledge graphs, they do not create a repository and are not applied to scholarly articles on dementia. Our proposed framework is different as it aims to store the information about the key findings of the RAs and not merely metadata knowledge of the scientific RAs. Therefore, our proposal in this paper is unique for creating a knowledge repository for knowledge discovery from dementia research findings. With our approach of rewriting the highlights of the RAs in the cognitive script, the knowledge extracted will be converted into Cypher statements to create a graph database. The query results from the graph database showing the outcomes of this research form a novel visualization of knowledge on dementia.

1.2 DSR approach for proposed model

DSR is a qualitative and multidisciplinary field research approach. In recent years, DSR is considered an exceptional research paradigm by information systems (IS) researchers for research concerned with the construction of socio-technical artefacts to solve IS problems and derive prescriptive design knowledge. The DSR approach is followed in this research to achieve the aims of the study to support the research community to store the existing findings in the domain of dementia disease as visually comprehensible knowledge graphs. Peffers (2007) developed a DSR methodology that was successfully embraced among the prominent IS development methods. The DSR methodology consists of the following steps:

1. Identification of the problem,
2. Definition of the solution objectives,
3. Design and development of the solution,
4. Demonstration of the solution,
5. Evaluation of the solution, and
6. Communication and discussion about the solution and evaluation.

We adopt the DSR methodology from Peffers (2007) for this research project with the framework for implementation as depicted in Fig. 1. The study started with a detailed description of the issues related to knowledge management of findings in the RAs. The problem is translated into a design requirement of solution based on cognitive script composition, NLP intelligent scripting and Cypher script generation. In the next step, an artefact is developed based on an illustrated example as identified in step 2 with a literature review. A system design is established based on NLP techniques and is applied as a prototype demonstrating

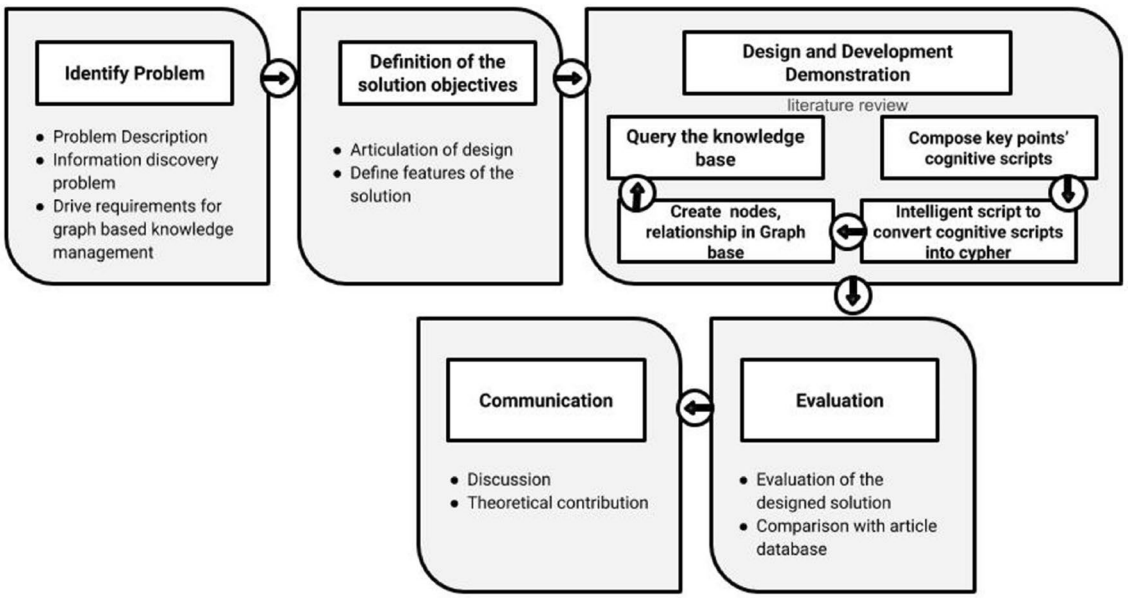

Fig. 1 Proposed model development using DSR approach

the applicability to dementia research as a case study. An evaluation has been carried out by performing visual navigation and searching of the graph database developed for this study. The result of the findings from the evaluation with an effective solution design is communicated through the DSR approach which also forms our theoretical contribution.

1.3 Computing model for developing a graph-based repository

This section describes our simple computing model using a validated DSR approach to develop the graph database to store knowledge from a collection of articles based on dementia and the approach for performing knowledge discovery and evaluation as a proof-of-concept. The creation of the computing model involved three phases and the evaluation, with each component given below following DSR methodology as described previously:

1. Article collection and construction of cognitive script,
2. Generation of Cypher statements, and
3. Creation of graph-based repository and visualization.

Figure 2 provides an overview of the process flow of our proposed model for developing the implementation framework including input and expected output. Firstly, the input is a set of unstructured text that is a collection of abstracts of scholarly articles on dementia, taken from the ScienceDirect database. The input is converted into a human readable cognitive script for the key findings of the RAs. Secondly, the intelligent Python program is applied to the cognitive script to generate the Cypher statements for storing this information in the graph-based repository. Afterwards, these output Cypher statements are executed to create entities with properties and their corresponding relationships in the repository. In the last phase, data from the repository are represented as a visual display and queried to perform knowledge discovery and evaluation. Figure 3 shows each step with an example.

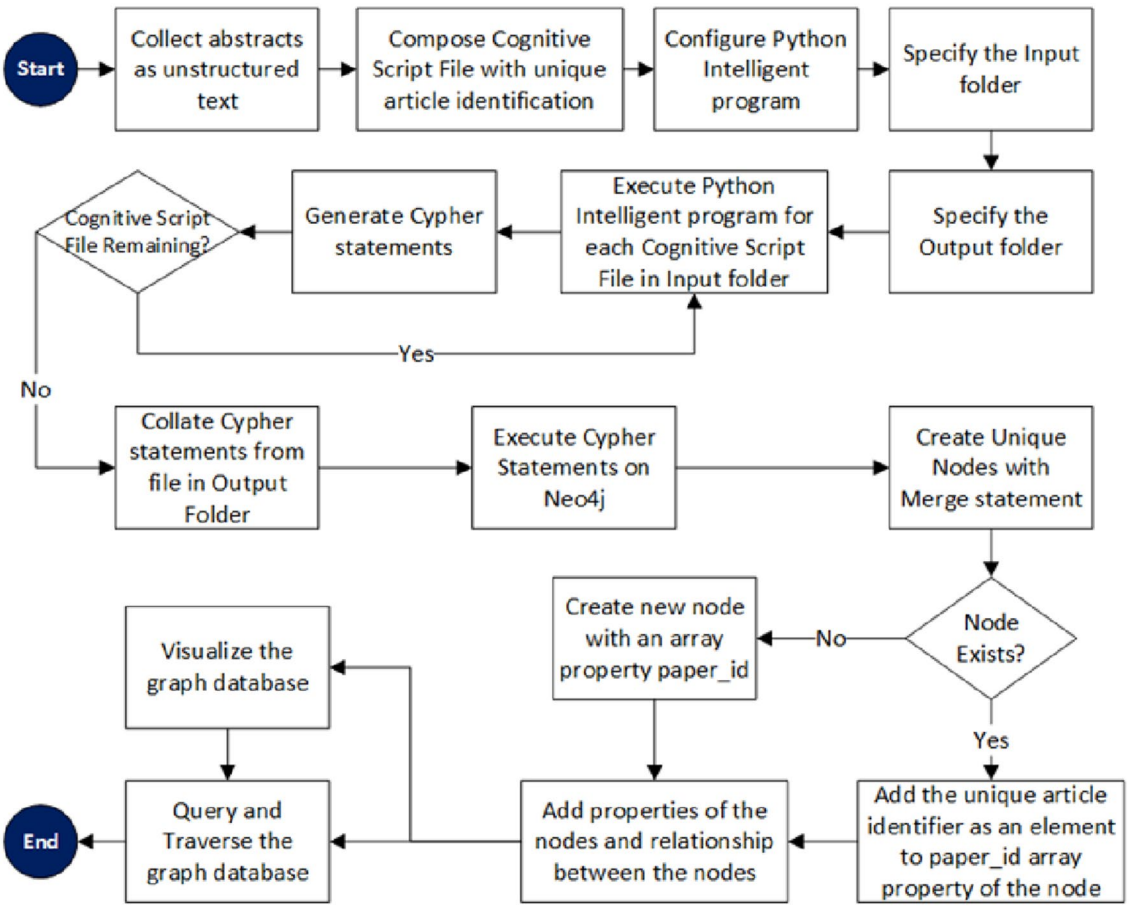

Fig. 2 Workflow of the Computing model using the cognitive script and graph-based infrastructures

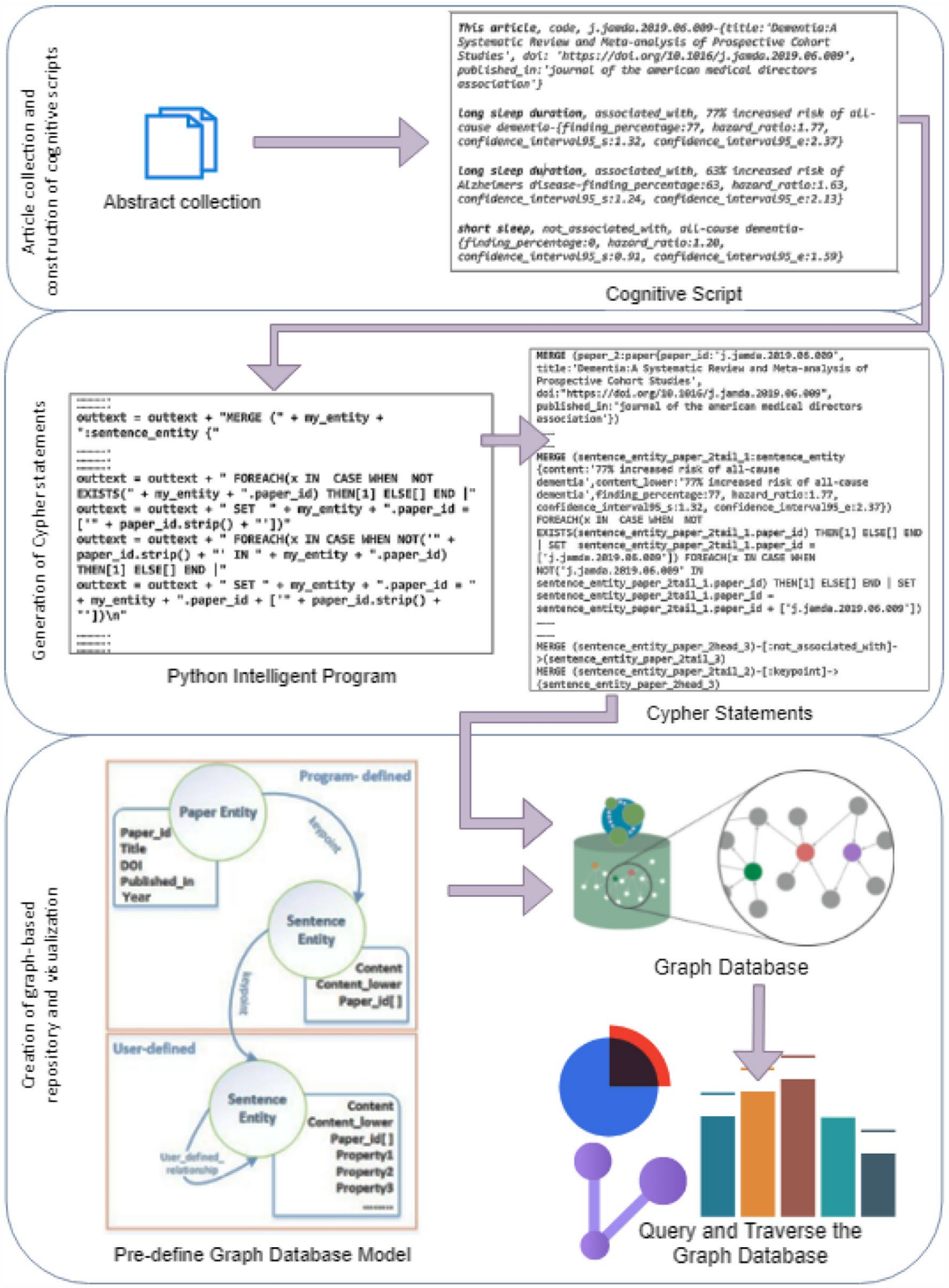

Fig. 3 Workflow of the Computing model using the cognitive script and graph-based infrastructures

2 Article collection and construction of cognitive scripts

To discover knowledge from RAs, the first step is to understand the organization of content as natural language text in the abstracts of RAs. Critical text analysis is performed on a collection of abstracts of scholarly articles on dementia from the ScienceDirect database. ScienceDirect is a database for science and medical journals to access 16 M articles and more than 39,000 e-books. This study performed a Systematic Literature review (SLR) to collect articles related to dementia and the risk factors of dementia. Search phrases like "Dementia","incident dementia", "risk factors of incident dementia" and "risk factors of incident dementia hazard ratio" were used to search and filter RAs. The search results were crowded; therefore, only the latest 100 articles were retrieved based on the year of publication. These retrieved articles were further narrowed down by critically reviewing the abstract, methodology for the number of participants and results for hazard ratio information. For experimental proposed only 10 articles were used as a collection to simplify the trial of the purposed computing model.

The study creates the dataset of the cognitive script from the sample research articles consisting of unstructured data. The cognitive script is a delimited text file consisting of a sequence of context-specific findings in each line separated by a comma. The key findings of each abstract of the research article collection are written in cognitive scripts to construct Cypher by following a few simple and systematic rules. The script starts with specifying paper id as a unique identifier for the RA for the key findings. The syntax is "start node, relationship name, end node" for each sentence, and "-" is used to define the attributes of each node which can be text-based or numeric-based. Figure 3 shows an example of cognitive scripts constructed for a key finding of an article.

3 Generation of Cypher statements

This study investigates a non-relational graph store, Neo4j, as a structured repository for the extracted knowledge. Graph store is a collection of machine-readable entities in the domain as nodes, and the relationships between entities to connect the nodes. It contains the properties of entities and the properties of the relationship between the entities as a key–value pair to describe detailed information. Traversing the nodes and relationships between the nodes in Neo4j is more efficient than other structured data sources (Robinson et al. 2013).

The output of step 4.1 is a semi-structured knowledge in the form of a cognitive script harvested from abstracts of RAs. This form of data is readable by both humans and computers. The cognitive script is then inputted to a Python-based intelligent script to generate Cypher statements. We provide Algorithm 1 that reads the cognitive script and generates the Cypher statements to create nodes, relationships and properties in the graph-based repository.

Code snippet of the Python program in Fig. 3 shows that a merge statement is used to merge the extracted knowledge from the cognitive script. The Merge Cypher statement avoids any duplication of the nodes. In our proposed approach, the Merge statement will create the unique nodes about a key point of an article and if that key point node already exists, adds article identification (paper id) as an element to the array property of the node and adds the relationship of the article to the existing node of the key point.

Figure 3 shows the output of the Python program with the input of the cognitive script. At first, nodes and their properties are created, then the head and tail of a connection are defined, and finally, the relationships are linked. Since establishing nodes and links with Cypher is complicated for humans, the Python script translates the human-readable sentence to Cypher queries. Thus, it provides a bridge for people to pass over this stage resulting in a human-readable format.

4 Creation of graph-based repository and visualization

We adopt a graph data structure to represent the nodes (called vertices) and the relationships between the nodes (called edges or arcs). The extracted structured knowledge is in the form of Cypher Query Language, aka Cypher which is a query language of Neo4j (Comito et al. 2004). The extracted knowledge is added to the existing repository of a graph database. It gets merged into the existing knowledge to extend the knowledge in the repository. This means that if there is a new knowledge, which can be a new node, a new relationship, or a new attribute that does not already exist in the repository, it gets created. The Cypher statements are executed in Neo4j to create the graph-based repository from the knowledge extracted from the text obtained from various RAs. These Cypher statements, consisting of clauses, keywords and expressions, are transformed into nodes, their properties and relationships, i.e. a graph-based repository data model to generate Cypher query as shown in Fig. 3.

There are two major entities in the data model of the graph-based repository: (i) paper-entity that records the information about the paper like paper title, DOI, journal name from cognitive scripts; and (ii) sentence entity that records the highlights of the paper like the content and links to the content of the same paper or other papers through a relationship. The graph repository is visually represented

by labelled schema drawing of graph structure data source with nodes and edges.

4.1 Case study

The representation of the graph-based repository of key findings of the collection of articles is displayed in Neo4j as nodes and edges of knowledge extracted from RAs. The graph is generated directly from the Cypher queries. The corresponding knowledge is shown by nodes and relationships in the graph; The article metadata knowledge is presented through the orange node and its properties. The associated findings of each article are shown through the blue nodes and their properties. The Cypher statements created 67 nodes stored permanently with an average of 4 properties each and 164 relationships in the Neo4j database based on the knowledge extracted from the key findings of each abstract of the collection of articles.

Cypher Query Match statement is used to retrieve all the information in the database in a visual graph form as: "MATCH (n)-[r]-(m) RETURN n,r,m". The graph visualization clearly shows the data and the relationship within the data. The graph can be traversed to find any information or answer any query. The use of coloured nodes helps to clearly distinguish or identify knowledge about the specific domain like metadata entities of articles that are represented by orange colour nodes or the key points entities that are represented by blue colour. The flexibility of creating a graph and extracting information from it provides advantages to make an advanced search.

4.2 Discussion and evaluation of the proposed model

In this paper, we proposed a DSR approach-based computing model that was developed and successfully implemented to import the key highlights of scholarly articles on dementia into a graph database in Neo4j. This data when queried appropriately can help to visualize the relations between the information published in RAs. Some of the visualizations created as a prototype evaluation include journal-wise publications, publications based on the number of participants, Study duration or highlighted information like hazard ratio (HR). We provide various Cypher queries and resulting visualization from the graph database developed from cognitive script to demonstrate the results achieved from the dementia RAs considered for this study.

Figure 4 depicts meta-analysis of publications by visually displaying the year-wise and journal-wise information. We used the Cypher queries on the graph database to extract the needed information for plotting each graph, i.e. MATCH (p:Paper) RETURN p.year, count(*) and MATCH (p:Paper) RETURN p. published in, count(*) (Fig. 5).

In a graph-based repository, it is possible to search based on the relationship between different nodes of information. For example, It is possible to find articles that have listed the risk of dementia and the factor that contributes to the risk of dementia by using the relationships like "keypoints"(of the article) and "risk associated with". The Cypher query used is MATCH (p:paper)-[:keypoint]->(f)-[:associated with]-(o:sentence entity) WHERE o.content = ~'.*risk of dementia.*' RETURN p,f,o

This graph-based repository is able to search on numeric values as well. For example to facilitate search phrases like "risk of dementia hazard ratio greater than 1.0", Cypher query return the required result MATCH (p:paper)-[r1]-(m)-[r2]-(f:sentence entity) WHERE (f.hazard ratio > 1) RETURN DISTINCT p,r1,m,r2,f. Another significant and similar filter is of range of participants to either find or restrict the RAs. In this repository, it is possible to search the RAs based on a range of the number of participants and the outcome of the search can be displayed in multiple formats. For example, to find RAs on the risk of dementia with the number of participants who underwent the research study lying between 2000 and 2900, the following query is created and our model gave the results successfully as shown in Fig. 6.

The evaluation supports the objective of the paper to facilitate the research community to link the discoveries with

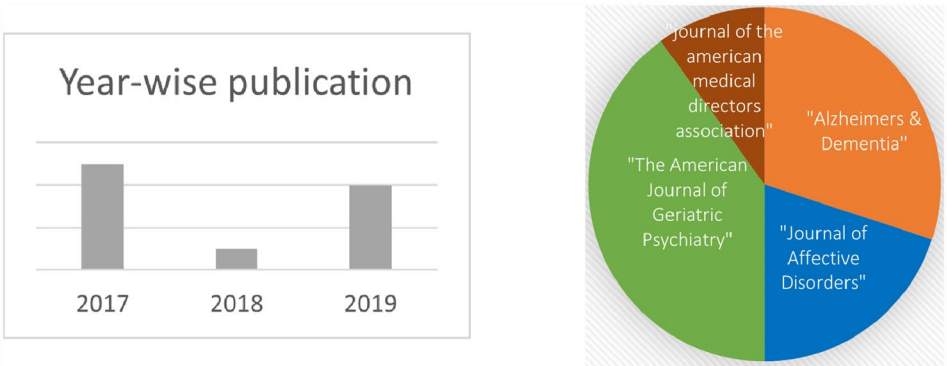

Fig. 4 Article Publications per Year and Journal

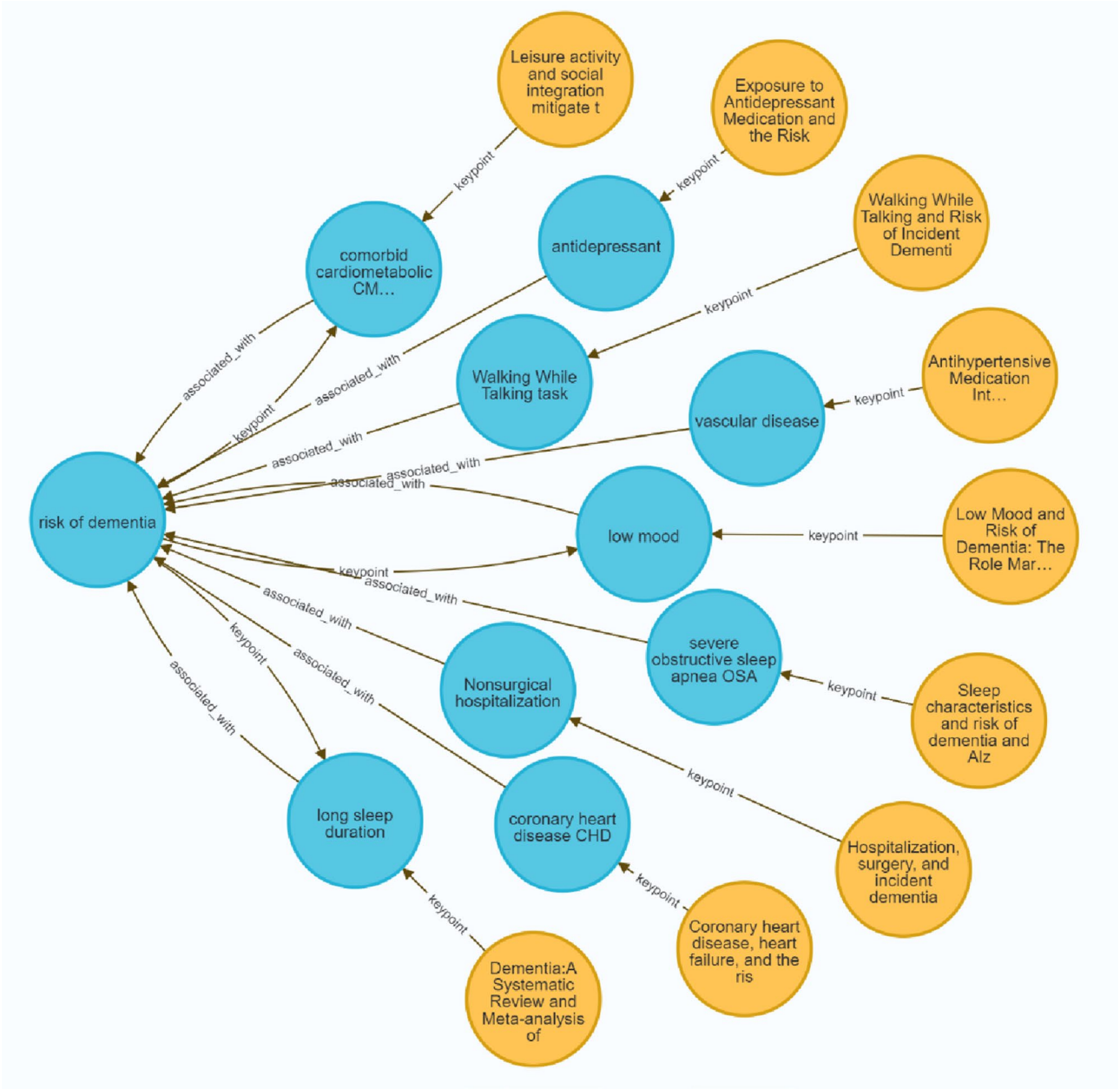

Fig. 5 Article publication and the risk factors of dementia

Fig. 6 Article publications and number of participants-based search

"p.title"	"p.number_of_participants"
"Leisure activity and social integration mitigate the risk of dementia related to cardiometabolic diseases: A population-based longitudinal study"	2648
"Low Mood and Risk of Dementia: The Role of Marital Status and Living Situation"	2599
"Hospitalization, surgery, and incident dementia"	2529

Table 1 Algorithm—Cypher query generation

Algorithm 1—Generating Cypher from cognitive script
1.**function generatecypherArticleNode (S)**
2.read the first line of the cognitive script
3.P_{id} = unique article identity = concatenation (prefix of P_{doi}, suffix of P_{doi}, rand number)
4.generate Cypher statement for article node with properties
5.append Cypher statement to O_{cypher}
6.**end function**
7.**function generatecypherKeypointNode(L)**
8.read and split the line
9.create **merge** Cypher statement of the node to create unique node only
10.if ((P_{id} NOT EXISTS as an Array property of the node) **then**
11.create P_{id} as an element of P_{id} Array property of the node
12.**else if** (P_{id} NOT EXISTS in P_{id} Array property of the node) **then**
13.add P_{id} as an element of P_{id} Array property of the node
14.**end if**
15.**end function**
16.**function createoutputFile(O_{cypher})**
17.create output file in the output folder
18.save O_{cypher} in the output file
19.**end function**
20.——————**Start**——————
21.O_{cypher} = null
22.**for each cognitive file in the input folder do**
23.S = read cognitive file content
24.O_{cypher} = generateArticleidNodeCypher(S)
25.**for each line L in S do**
26.O_{cypher} = O_{cypher} + generatecypherKeypointNode(L)
27.O_{cypher} = O_{cypher} + generatecypher for properties of the node
28.O_{cypher} = O_{cypher} + generatecypher for relationship
29.**end for**
30.createoutputFile(O_{cypher})
31.**end for**
32.——————**End**——————

the current knowledge in the domain and possibly identify gaps that would facilitate new research opportunities.

The study follows the DSR methodology as defined earlier. The study begins with Step 1—problem identification related to knowledge management of highlights of RAs provided in Sect. 1. Step 2 is given in Sects. 2 and 3, which identify the gap with the literature review of existing knowledge to articulate the design of the computing model as a solution. Section 4 defines Step 3 of design, development and demonstration of the computing model based on cognitive script composition, NLP scripting and cypher script generation. Next DSR methodology step 5 of the evaluation of the computing model is given in Sect. 5. These DSR methodology steps of designing, developing, demonstrating and evaluating the computing model communicated through a theoretical contribution of this paper as the last step.

5 Conclusion and future works

To collect and record information from scholarly articles for knowledge discovery by humans and machines is a complex process. This paper has tried to address this by automating the process with a computing model. We proposed the model using a cognitive script for extracting information from RAs to store the highlights of the articles as a graph-based repository. Automatic creation of the repository is challenging which was attempted in this study. With our proposed novel framework using DSR methodology, we stored unstructured knowledge from RAs by composing semi-structured cognitive scripts, generating Cypher statements and integrating with existing structured data in a graph-based repository. The framework is evaluated on a set of abstracts obtained from research papers on dementia. We have demonstrated that the semi-automatic creation of a graph-based repository from knowledge extracted from scholarly articles on dementia can support researchers to have a better perception of the research findings in the domain. The transformation from the human readable cognitive script to Cypher query language facilitated comprehension and effectively allowed human involvement in the research community. The knowledge graph was used to visualize the relationships between the knowledge of the same article and relevant information from the other articles from the domain of dementia. Related knowledge was integrated into a more comprehensible form based on the knowledge graph either as properties or relationships of the entities. Thus, the discovery of the findings in the dementia related RAs is improved through the graph-based storage and retrieval of information.

The main limitation is the article collection used for demonstrating and evaluating our proposed model as a proof-of-concept in this paper. However, the model can be easily implemented to process a larger collection related to dementia. Furthermore, the DSR methodology (Miah et al. 2019a; Miah 2008) used in this study as an iterative process facilitates the development of the model in this work with an article collection as only a representative of a bigger population achieving continuous improvements in future iterations. Besides, the impact of the understanding of the narrative format and results from the database is not directly measurable. In future, we are intending to improve the model so that it will be able to read simple sentences directly and perform semantic division. This would facilitate automatically creating the cognitive scripts to convert the article content into the graph-based knowledge repository. Furthermore, the next step of our study will focus on automatic extraction of highlights directly from the scholarly articles by using

pre-trained models like BERT or GPT3 instead of developing cognitive text. Future work could be devoted towards enhancing the proposed model by generating the links with the existing ontology and architectural choices such as Energy-Aware Scheduling Strategy by Comito et al. (2011). Future research directions would also consider existing design research methodological works (Genemo et al. 2015) such as for big data oriented goverance solutions (e.g. Miah et al. 2021; Miah et al. 2019b) to ensure efficient routing, resource allocation and workload management.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

ABS (2021) Causes of death, Australia, https://www.abs.gov.au/statistics/health/causes-death/causes-death-australia/2020

AIHW(2020) Dementia, https://www.aihw.gov.au/reports/ 475 australias-health/dementia/

Aslam M, Aljohani N, Abbasi R, Lytras M, Kabir M (2017) A generic framework for adding semantics to digital libraries

Balaida A, Rozana M, Hikmia S, Memon J (2016) Knowledge maps: a systematic literature review and directions for future research. Int J Inf Manage 36:451–475

Bayram U, Roy R, Assalil A, BenHiba L (2021) The unknown knowns: a graph-based approach for temporal COVID-19 literature mining. Online Inf Rev 45(4):pp 687–708. Doi: https://doi.org/10.1108/OIR-12-2020-0562

Bird S, Klein E, Loper E (2009) Natural Language Processing with Python. O'Reilly Media

Comito C, Talia D (2004) GDIS: a service-based architecture for data integration on grids. In: Meersman R, Tari Z, Corsaro A (eds) On the move to meaningful internet systems 2004: OTM 2004 workshops. OTM 2004. Lecture Notes in Computer Science, vol 3292. Springer, Berlin. Doi: https://doi.org/10.1007/978-3-540-30470-8_27

Comito C, Patarin S, Talia D (2007) PARIS: A Peer-to-Peer architecture for large-scale semantic data integration. In: Moro G, Bergamaschi S, Joseph S, Morin JH, Ouksel AM (eds) Databases, information systems, and peer-to-peer computing. DBISP2P 2006, DBISP2P 2005. Lecture Notes in Computer Science, vol 4125. Springer, Berlin. Doi: https://doi.org/10.1007/978-3-540-71661-7_15

Comito C, Falcone D, Talia D, Trunfio P(2011) Energy efficient task allocation over mobile networks. In: 2011 IEEE ninth international conference on dependable, autonomic and secure computing, pp 380–387. Doi: https://doi.org/10.1109/DASC.2011.80.

Comito C, ForestieroA, Pizzuti C (2019) Word embedding based clustering to detect topics in social media. In: 2019 IEEE/WIC/acm international conference on web intelligence (WI), pp 192–199

Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805

Gardner M, Grus J,. Neumann M, Tafjord O, Dasigi P, Liu N, Petersm P, Schmitz M, Zettlemoyer L (2018) AllenNLP: a deep semantic natural language processing platform. Allen Institute for Artificial Intelligence

Gangemi A (2013) A comparison of knowledge extraction tools for the semantic web. In: The semantic web: semantics and big data. Springer Berlin, pp. 351–366.

Gangemi A, Presutti V, Recupero D, Nuzzolese A, Draicchio F, Mongiov M (2017) Semantic web machine reading with fred. Semantic Web 8:873–893

Genemo H, Miah SJ, McAndrew A (2015) A design science research methodology for developing a computer-aided assessment approach using method marking concept. Educ Inf Tech 21:1769–1784

Google (2021) Google scholar, https://scholar.google.com/intl/en/scholar/about.html

Hansen B, Kautz K (2004) Knowledge mapping: a technique for identifying knowledge flows in software organisations

Iorio F (2015) Cognitive autonomy and methodological individualism : the interpretative foundations of social life. Springer, Cham

Kaur A, Chopra D (2016) Comparison of text mining tools. In: 5th International conference on reliability, infocom technologies and optimization (Trends and Future Directions) (ICRITO), pp 186–192

Kertkeidkachorn N, Ichise R (2018) Automatic knowledge graph creation framework from natural language text. IEICE Trans Inf Syst, pp 90–98

Lal M (2015) Neo4j Graph Data Model. Packt

Lam H, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, Liu N, Crasto C, Morse T, Stephens S, Cheung K (2007) Alzpharm: integration of neurodegeneration data using rdf, BMC Bioinformatics 8

Maksutov AA, Zamyatovskiy VI, Vyunnikov VN, Kutuzov AV (2020). Knowledge base collecting using natural language processing algorithms. In: 2020 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus), Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2020 IEEE Conference Of, pp 405–407. Doi: https://doi.org/10.1109/EIConRus49466.2020.9039303

Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The stanford corenlp natural language processing toolkit. Assoc Comput Linguist

Markus N, Hartung M, Ngomo A, Rahm E (2015) A survey of current link discovery frameworks. Semantic Web 8:419–436

Miah SJ (2008) An ontology based design environment for rural decision support, Unpublished. Griffith Business School, Griffith University, Australia PhD Thesis

Miah SJ, Gammack JG, McKay J (2019a) A metadesign theory for tailorable decision support. J Assoc Inf Syst 20(5):570–603

Miah SJ, Vu HQ, Gammack J (2019b) A big-data analytics method for capturing visitor activities and flows: the case of an Island Country. Inf Tech Manag 20(4):203–221

Miah SJ, Camilleri E, Vu HQ (2021) Big data in healthcare research: a survey study. J Comput Inf Syst 62(3):480–492

Neubig G,. Dyer C,. Goldberg Y,Matthews A, Ammar W, Anastasopoulos A, Ballesteros M, Chiang D,. Clothiaux D, Cohn T (2017) Dynet: The dynamic neural network toolkit: computation and language and mathematical software

Paulheim H (2016) Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 0, vol 1

Peffers K, Tuunanen T, Rothenberger M, Chatterjee S (2007) A design science research methodology for information systems research. J Manag Inf Syst 24:45–77

Qi P, Zhang Y, Zhang Y, Bolton J, Manning C (2020) Stanza: a python natural language processing toolkit for many human languages. ACL2020 System Demonstration

Robinson I, Webber J, Eifrem E (2013) Graph satabases, O'Reilly Media

Semantic Scholar (2021) A new and improved semantic scholar API,https://medium.com/ai2-blog/a-new-and-improved-semantic-scholar-api-8dd6329972bc

Tiwana A (1999) Knowledge management toolkit, the amrit tiwana knowledge management toolkit. Prentice Hall PTR

White D (2002) Knowledge mapping and management, IRM Press, London

Zhu Q, Luo J (2021) Generative pre-trained transformer for design concept generation: an exploration. ArXiv, abs/2111.08489. https://doi.org/10.48550/arXiv.2111.08489