Actionable Knowledge Extraction Framework for COVID-19

Mohammad Masum
Analytics and Data Science Institute
Kennesaw State University
Kennesaw, USA
mmasum@students.kennesaw.edu

Hossain Shahriar
Department of Information Technology
Kennesaw State University
Marietta, USA
hshahria@kennesaw.edu

Hisham M. Haddad
Department of Computer Science
Kennesaw State University
Marietta, USA
hhaddad@kennesaw.edu

Sheikh Ahamed
Department of Computer Science
Marquette University, USA
sheikh.ahamed@marquette.edu

Sweta Sneha
Department of Information Systems
Kennesaw State University, USA
ssneha@kennesaw.edu

Mohammad Rahman
Department of Electrical and Computer Engineering
Florida International University, USA
marahman@fiu.edu

Alfredo Cuzzocrea
iDEA lab
University of Calabria, Italy
alfredo.cuzzorea@unical.it

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378398

***Abstract*— In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19) containing over 51,000 scholarly articles, including over 40,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. Medical professional including physicians frequently seek answers to specific questions to improve guidelines and decisions. The huge resource of medical literature is important sources to generate new insights that can help medical communities to provide relevant knowledge and overall fight against the infectious disease. There are ongoing attempts to develop intelligent systems to automatically extract relevant knowledge from many unstructured documents. In this paper, we propose an efficient question answering framework based on automatically analyzing thousands of articles to generate both long text answers (sections/ paragraphs) in response to the questions that are posed by medical communities. In the process of developing the framework, we explored natural language processing techniques like query expansion, data preprocessing, and vector space models early. We show the initial results of an example query answering for the incubation period.**

Keywords*— *Coronavirus, COVID-19, Query expansion, TF-IDF, Knowledge extraction, Natural Language Processing, Text mining

I. Introduction

An outbreak of pneumonic disease has been spreading since December 2019 from Wuhan, the largest metropolitan area in Hubei province of China [5]. Since then the world has been experiencing an outbreak of pneumonia caused by a novel coronavirus known as severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) or COVID-19 early 2020 [1]. The epidemic was declared as a Public Health Emergency of International Concern on 30 January 2020 and recognized as a pandemic on 11 March 2020 by the World Health Organization [2].

The respiratory disease, spreading from person-to-person, has caused the pandemic in about a quarter of a year, affecting 230 nations worldwide. The infectious disease has caused 1,779,842 confirmed cases and 108,779 confirmed deaths by April 11, 2020, fundamentally affected the USA, Italy, Spain, and France [6].

In response to the pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19) containing over 51,000 scholarly articles, including over 40,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses [3]. Medical professional including physicians frequently seek answers to specific questions to improve guidelines and decisions. The huge resource of medical literature is important sources to generate new insights that can help medical communities to provide relevant knowledge and overall fight against the infectious disease. Consequently, there are ongoing efforts to develop intelligent systems to automatically extract relevant knowledge from many unstructured documents.

Manually relevant knowledge extraction from the large volume of articles is labor-intensive, unscalable and challenging [4, 15]. Consequently, there have been several attempts to develop intelligent systems to automatically extract relevant knowledge from many unstructured documents. Some approaches focus on the idea of extracting knowledge and supporting big data analytics over networked data [17, 18]. In this paper, we propose to develop an efficient question answering framework based on automatically by analyzing thousands of articles to generate both long text answers (sections/ paragraphs) in response to the questions that are posed by medical communities. The automatic knowledge extraction system can generate meaningful information from the thousands of articles to assist the healthcare professionals and government officials in decision making process. In the process of developing the framework, we explored natural language processing techniques like

query expansion, data preprocessing, and vector space models early.

Our contributions of this paper are two folds: 1) The question- answering framework for knowledge extraction can extract most relevant long-text answers in terms of sections/paragraphs from the articles related to a specific question or query. 2) In the 2nd phase, the framework provides a similarity network of relevant sections generated from given articles. The network will provide a broader point of view about a specific query. In this process, a healthcare professional will not only be informed from one related article but also from other articles that discussed about the topic.

This paper shows a case study of the application of the proposed framework about incubation period for the disease in humans [3], which is still an active research area among professionals.

The paper is organized as follows. Section II discusses related works. Section III describes the proposed methodology. Section IV provides an example dataset analysis and knowledge extraction based on the proposed method. Finally, Section V concludes the paper.

II. Related Works

An effective knowledge extraction system can help accelerate healthcare professionals in decision making process if implemented correctly. Existing methods for knowledge extraction can be categorized into three: (1) keyword matching, (2) grammar analysis, and (3) rule based regular expression matching methods. Knowledge can be derived from keyword matching technique by matching user-defined keywords to text in given documents [10, 11]. Texts in a document can be tokenized by a single space or a new line to match with the user-defined keywords. It considers that all terms in the document are independent of each other. The performance of this approach depends on the provided keywords.

Performance of the knowledge extraction system frequently depends on detecting relationship between words, that have a higher probability of occurring together or have a close association. The association between the words is extracted and leveraged by grammar analysis method [7, 13, 14]. This approach is limited to finding the relation between verb-adjective or a noun-verb which follows

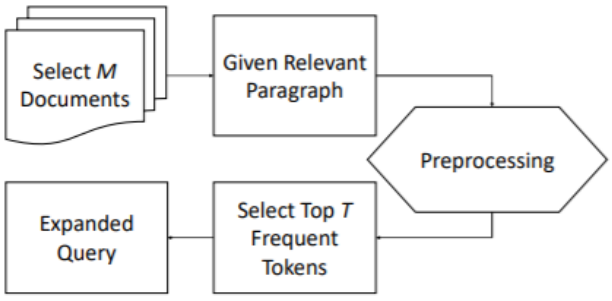

Figure 1: The Process of Generating Expanded Query

grammar rules, but two closely related nouns or verbs cannot be related, for instance, "price" and "payment" are nouns, which are closely related in a business domain, but their relationship cannot be defined by grammar analysis. To obtain the relationship between two closely related words, rule-based regular expression matching systems have been proposed [14], where a set of rules is pre-defined by matching regular expression or searching with multiple keywords. However, this approach confines only a set of documents which follow the designed rules and involves domain experts to define rules, which can be expensive and hard to generalize the solutions.

Different natural language processing models are widely used for knowledge extraction from documents [16]. An efficient framework for knowledge extraction system was proposed that can extract efficient information from thousands of unstructured OCR contract documents [4]. The framework used rule-based methods for converting unstructured OCR documents into hierarchical structed JSON format and then vector space model was utilized for retrieving ranked relevant information from the documents. A knowledge extraction system was designed that can automatically identify important topics related to IoT (internet of things) from a corpus of 4500 conference and journal articles [8]. The extracted important topics then used for building a semantic Web of things. Automatic knowledge extraction utilizing Latent Dirichlet Allocation was applied to find relevant publications in a cluster of manufacturing research publications [9].

In this paper, we propose a framework where the relationship between the words is achieved by using Query expansion (QE) technique and knowledge extraction with the user defined query is achieved by a vector space model and hierarchical document analysis.

III. Methodology

A. Proposed Method

We developed an automatic knowledge extraction framework to extract the most relevant section of interest from a corpus of COVID-19 related research articles. The framework performed several steps to retrieve most relevant section given a specific given query. Retrieving the information, some steps were performed:

1. *Query Expansion:* Using short query for information retrieval leads to term mismatch information retrieval problem since the query term may lack quantity of enough words. Query expansion technique overcomes this limitation by adding new tokens (words) to the existing terms and generate expanded query. Local analysis is one of the QE techniques where the query is expanded by leveraging top ranked relevant retrieved section(s) by the given query. Relevance feedback is another QE method that utilize relevant documents provided by users [7]. Fig. 1 illustrates process of generating expanded query. *M* number of documents that are related to a given query (key words) are selected at first. Relevant paragraphs to the query are carefully

chosen from each of the selected documents by subject related experts. Following, the paragraphs are preprocessed (discussed next subsection) and new tokes (terms) are added by leveraging *n*-grams to the existing total tokens in all the paragraphs. Finally, top *T* frequent words are selected that is used as expanded query.

2. *Data Preprocessing:*
    a. *Case Folding:* In this step, all text in the documents are converted into small letter like the phrase "Data Mining" converted into "data mining".
    b. *Stop words removing:* Stop words are the words such as a, an, the, that, etc. in a document that contain trivial information. Stop words should be removed in the process of model building.
    c. *Special character removing:* There are some special characters in the documents like "?", "#", "!" etc. which should be removed.
    d. *Lemmatization:* Different forms of a words (like studies, studying) are may be used in the documents for grammatical reasons. Lemmatization is used to reduce inflectional forms of words to a common base form (studies, studying → study).
    e. *N-grams:* N-grams considers N consecutive words instead of a single word. For instance, 2-grams of the phrase: "Transmission dynamics of the virus" → "Transmission dynamics", "dynamics of", "of the", and "the virus". It can capture the context of text in a document.
3. *Transformation- Vector Space Model using TF-IDF:* Documents are at first transformed into Bag of Words (BoW) by splitting each of the sentences into words. Later, the BoW are transferred into vector spaces by term frequency-inverse document frequency (TF-IDF) method. Term frequency (TF) is defined in a document by taking the ratio between total occurrence of a term (word) and total number of words in the document. On the other hand, inverse document frequency (IDF) measures importance or weight of a term to a document in collection of corpora [4]. Equations 1, 2, and 3 show the definitions of TF, IDF, and TF-IDF in mathematical form.

$$\boldsymbol{tf} = \frac{frequency\ of\ a\ term\ \boldsymbol{t}\ in\ a\ document}{total\ terms\ in\ the\ document} \quad \textbf{(1)}$$

$$\boldsymbol{idf} = \frac{total\ numer\ of\ docuemnts}{number\ of\ docuements\ with\ the\ term\ \boldsymbol{t}} \quad \textbf{(2)}$$

$$\boldsymbol{tf-idf} = tf \times idf \quad \textbf{(3)}$$

4. *Similarity calculation:* Cosine similarity is one of the methods to calculate similarity between pairs of documents. It measures the cosine of the angle between two vectors where each of the vectors are numerical presentation of documents (sentences/ sections/ paragraphs) derived by using vector space models. Equation 3 shows the mathematical definition of cosine similarity where A and B are two documents.

$$similarity = Cos(\theta) = \frac{\boldsymbol{A}.\boldsymbol{B}}{||\boldsymbol{A}||||\boldsymbol{B}||} \quad \textbf{(4)}$$

5. *Information Extraction:* In this step, the framework retrieves the relevant information to a query by utilizing the similarity score between pairs of queries and all the documents.
6. *Similarity Network:* A weighted similarity network was developed in the 2nd phase of the project to gain more actionable insights from the research articles. In this step, we will utilize the output of stage 1.

IV. Experiment and Results

A. Dataset Specification

In response to the pandemic, the White House and a coalition of leading research groups have prepared and made available of the COVID-19 Open Research Dataset (CORD-19) containing over 51,000 scholarly articles, including over 40,000 with full text papers on COVID-19, SARS-CoV-2, and related coronaviruses [3]. In this project, only research articles from biorxiv_medrxiv folder consisting of 803 full text papers are analyzed from these broad corpora. The given papers are in json format which requires conversion to structured format for further study.

Fig 2 shows a snippet of a sample paper in JSON format. Each of the papers consists of 7 sections: paper_id, metadata, abstract, body_text, bib_entries, ref_entries, and back_matter. We only extracted article title (from paper_id), abstract, and body text from each of the JSON formatted paper for this study. Fig. 3 illustrates extraction of title, abstract, and body text from a JSON formatted paper. Later, the extracted texts are concatenated to form a full paper text.


```
{'title': 'Nucleotide Analogues as Inhibitors of Viral Polymerases',
 'authors': [{'first': 'Jingyue',
   'middle': [],
   'last': 'Ju',
   'suffix': '',
   'affiliation': {'laboratory': '',
    'institution': 'Columbia University',
    'location': {'postCode': '10027',
     'settlement': 'New York',
     'region': 'NY'}},
```


Figure 2: Text in JSON format

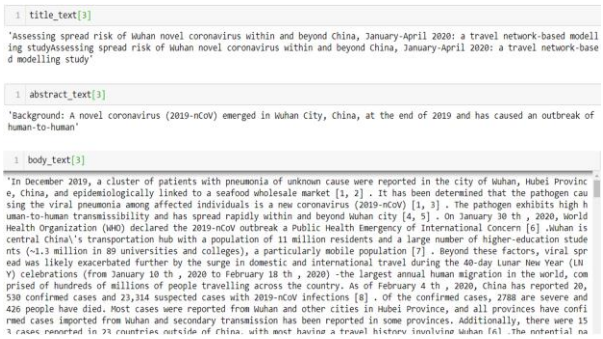

Figure 3: Extraction of title, abstract and body text

B. Experiments

The process of retrieving most relevant section or paragraphs from the articles is illustrated in Fig. 4. For instance, the user feed the input with a query term "incubation period". Relevance feedback query expansion technique was then applied on the query and expanded query were generated for further study [6]. We manually selected 5 articles that contains information related to incubation period of corona virus. Later, important paragraphs/ sections were selected and preprocessed to new tokens. Finally, we opted to use most 50 frequent tokens that includes both words from the selected sections and bigrams of the terms. Table 1 shows the most frequent 50 tokens that were used as expanded query for term "incubation period". "incubation period", "quarantine period", "symptom onset", "period days", and "fatality rate" are the included in the expanded query as bigrams. Other frequent words like "days", "isolation", "infect", "time", "disease", "median", "mean", "range", "expose", "study", and "estimate" are expected to be in the same section that discussed incubation period.

The knowledge extraction framework searches for most relevant articles at first and then hierarchically investigates the most relevant (sub)section of each articles and finally return a ranked based relevant section to the user. Fig. 5 shows a similarity network of the query term "incubation period" by using a weighted graph where the weights are the similarity score. The most relevant articles are id-789, id-227, id-340, id-375, and id-621 with similarity scores of 0.17, 0.149, 0.139, 0.133, and 0.123, respectively. Since the framework hierarchically investigate the relevant section from the most relevant article, the similarity network shows the hierarchical relevant information as well. The 2nd layer displays the similarity between query and sections contained in the retrieved most relevant article (id: 789). The article (id -789) contains 8 sections in total where section 7 is the most related to the query with a similarity score of 0.206. Section 4 and 3 are the 2nd and 3rd most relevant sections with scores of 0.157, and 0.152, respectively.

Table 2 displays the retrieved most relevant three section from the most relevant article (id - 789) to the query term "incubation period". The "id" is the index of section of the most relevant article, similarity score measures the correlation between the given query and the sections of the article and the relevant section shows the extracted relevant information from the documents in terms of section (full section or snippet of a section). The most relevant article (id - 789) to the query is titled "Transmission of corona virus disease 2019 during the incubation period may lead to a quarantine loophole". The information in the sections may assist the healthcare professionals to model the current and future scale of the infectious diseases and assessing disease control strategies like deciding how many quarantine days are required for persons who may have been exposed to the virus. For instance, the most relevant section discusses infection during the incubation period time of COVID-19. Section 4 in the article mentions the mean incubation period of COVID-19 and section 3 examines transmission during incubation period. Therefore, the section contains many significant information related to the incubation period of novel coronavirus that can be important to the medical professionals to take decisions.

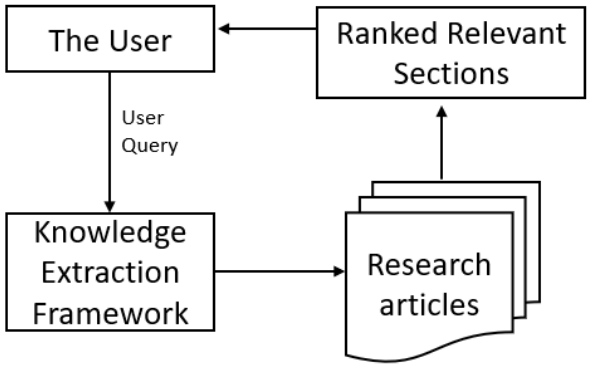

Figure 4: Process of Knowledge Extraction

Table 1: Frequent Terms for Generating Expanded Query

Terms	frequency	terms	frequency	terms	frequency
Case	26	onset	6	adults	4
Days	24	mean	6	longer	4
period	23	quarantine period	6	model	4
incubation	22	report	5	wuhan	4
incubation period	16	study	5	history	4
covid-19	11	hospitalize	5	symptom	4
quarantine	10	result	5	day	4
infect	10	periods	5	use	4

Terms	frequency	terms	frequency	terms	frequency
epidemic	10	data	5	casualties	4
time	8	estimate	5	symptom onset	4
individuals	7	fatality	5	fatality rate	4
disease	6	rate	5	outbreak	3
isolation	6	number	5	base	3
median	6	expose	5	therefore	3
range	6	period days	5	population	3
transmission	6	coronavirus	4	7-day	3
infection	6	include	4		

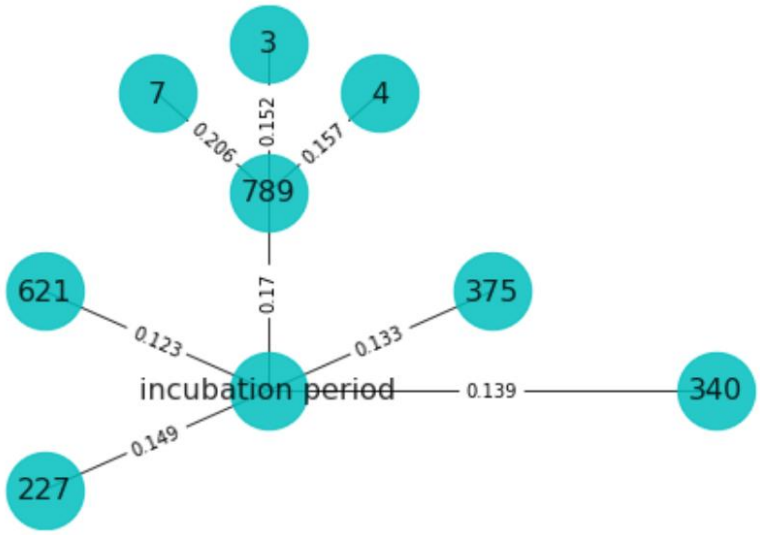

Figure 5: Similarity Network for Query term "incubation period"

Table 2: Relevant Sections and Similarity Score to the Query "Incubation Period"

Id	Similarity Score	Relevant Section (or a snippet of the section)
7	0.206	If the asymptomatic transmission during the incubation period contributes to a significant number of infec tion cases, a loophole of quarantine will ensue. In the present study, we aimed to examine whether and wh en the COVID-19 cases during their incubation period infect others, in order to offer clues to containment measures to mitigate the spread of infection. Accurate estimation of the incubation period is essential to cl arify whether the cases who are in the incubation period are a potential source of infection. The exposure t ime of the majority cases in Wuhan is ambiguous and therefore cannot be used to estimate the incubation period. By examining the cases who were confirmed to be infected with SARS-CoV-2 after a short visit t o Wuhan or having a short contact history with a confirmed case, the incubation period could be accuratel y estimated based on their precise exposure time.
4	0.157	The estimated mean incubation period for COVID-19 was 4.9 days (95% confidence interval [CI], 4.4 to 5.4) days, ranging from 0.8 to 11.1 days (2.5th to 97.5th percentile). The observed mean and standard dev iation (SD) of serial interval was 4.1±3.3 days, with the 2.5th and 97.5th percentiles at -1 and 13 days. Th e infectious curve showed that in 73.0% of the secondary cases, their date of getting infected was before s ymptom onset of the first-generation cases, particularly in the last three days of the incubation period.
3	0.152	We collected data on demographic characteristics, exposure history, and symptom onset day of the confir med cases, which had been announced by the Chinese local authorities. We evaluated the potential of tran smission during the incubation period in 50 infection clusters, including 124 cases. All the secondary case s had a history of contact with their first-generation cases prior to symptom onset.

V. Conclusion

The world is experiencing an outbreak of pneumonia caused by a novel coronavirus. The huge resource of medical literature related to coronavirus contain valuable information that can provide answers to specific questions posed by medical research communities. Health care professionals can improve their policies by efficiently analyzing and extracting coronavirus related specific knowledge from many published articles and overall fight against the infectious disease. We developed an actionable knowledge extraction framework that can automatically extract relevant information in terms of sections/ paragraphs related to a specific query. The framework leverages state-of-the-art natural language processing techniques in the process of model building. The framework also provides a similarity network generated from given articles that offer a broader point of view about a specific query. Therefore, a healthcare professional will not only be informed from one related article but also from other articles that discussed about the topic.

References

[1] Estrada, E. (2020). Topological Analysis of SARS CoV-2 Main Protease. bioRxiv.

[2] Ortea, I., & Bock, J. O. (2020). Re-analysis of SARS-CoV-2 infected host cell proteomics time-course data by impact pathway analysis and network analysis. A potential link with inflammatory response. BioRxiv.

[3 CORD-19 Challenge, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[4] Mohammad, M., Kosaraju, S., Bayramoglu, T., Modgil, G., & Kang, M. (2018, October). Automatic knowledge extraction from OCR documents using hierarchical document analysis. In Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems (pp. 189-194).

[5] Huang, Y., Tu, M., Wang, S., Chen, S., Zhou, W., Chen, D., & Huang, Q. (2020). Clinical characteristics of laboratory confirmed positive cases of SARS-CoV-2 infection in Wuhan, China: A retrospective single center analysis. Travel Medicine and Infectious Disease.

[6] https://www.worldometers.info/coronavirus/

[7] Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002, May). Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web (pp. 325-332).

[8] Noura, M., Gyrard, A., Heil, S., & Gaedke, M. (2019). Automatic knowledge extraction to build semantic web of things applications. IEEE Internet of Things Journal, 6(5), 8447-8454.

[9] Boonyasopon, P., Riel, A., Uys, W., Louw, L., Tichkiewitch, S., & Du Preez, N. (2011). Automatic knowledge extraction from manufacturing research publications. CIRP annals, 60(1), 477-480.

[10] Ahsan Mahmood, Hikmat Ullah Khan, Zahoor-Ur-Rehman, and Wahab Khan. 2018. Query based information retrieval and knowledge extraction using Hadith datasets. In Proceedings - 2017 13th International Conference on Emerging Technologies, ICET2017. https://doi.org/10.1109/ICET.2017.8281714

[11] Wenhao Zhu, Laihu Luo, Chaoyou Ju, and Bofeng Zhang. 2012. Cross language information extraction for digitized textbooks of specific domains. In Proceedings - 2012 IEEE 12th International Conference on Computer and Information Technology, CIT 2012. https://doi.org/10.1109/CIT.2012.226

[12] S. T. Zuhori, M. A. Zaman, and F. Mahmud. 2017. Ontological knowledge extraction from natural language text. In 2017 20th International Conference of Computer and Information Technology (ICCIT). 1–6. https://doi.org/10.1109/ICCITECHN. 2017.8281776

[13] R. Upadhyay and A. Fujii. 2016. Semantic knowledge extraction from research documents. In 2016 Federated Conference on Computer Science and Information Systems (FedCSIS). 439–445.

[14] Cuzzocrea, A., De Maio, C., Fenza, G., Loia, V., & Parente, M. (2016, April). OLAP analysis of multidimensional tweet streams for supporting advanced analytics. In Proceedings of the 31st Annual ACM Symposium on Applied Computing (pp. 992-999).

[15] Cuzzocrea, A. (2006). Combining multidimensional user models and knowledge representation and management techniques for making web services knowledge-aware. Web Intelligence and Agent Systems: An international journal, 4(3), 289-312.

[16] Liu, L., Priestley, J. L., Zhou, Y., Ray, H. E., & Han, M. (2019, December). A2Text-Net: A Novel Deep Neural Network for Sarcasm Detection. In 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI) (pp. 118-126). IEEE.

[17] Georgios Chatzimilioudis, Alfredo Cuzzocrea, Dimitrios Gunopulos, Nikos Mamoulis, "A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement". J. Comput. Syst. Sci. 79(3): 349-368 (2013)

[18] Alfredo Cuzzocrea, Il-Yeol Song, "Big Graph Analytics: The State of the Art and Future Research Agenda". DOLAP 2014: 99-101