Building Knowledge Graphs based on Binary Associations between Research Topics using Apriori

Reem Q. Al Fayez
Computer Information Systems department
University of Jordan, Amman, Jordan

Heba Saadeh
Computer Science department
University of Jordan, Amman, Jordan

Samar Saleh
Computer Information Systems department
University of Jordan, Amman, Jordan

Baraa Abu Alrub
Computer Science department
University of Jordan, Amman, Jordan

ABSTRACT

Academic and scientific research is very important in developing communities. Interdisciplinary research is thriving and it became the focus of many organizations in academia and industry. Scopus is one of the largest databases of peer-reviewed content. It is a very popular platform for researchers to visit when doing research. Scopus have developed the All Subject Journal Classification (ASJC) hierarchy to denote the research topics covered in their database. The relations between the subjects is not represented by the ASJC. It is represented in a tree like structure of topics and subtopics of research. Journals indexed in Scopus state research topics that they publish in using ASJC topics. The combination of such research topics can be used for discovering associations between research topics. Using Apriori for frequent patterns discovery of co-occurring ASJC topics can build a knowledge graph of associations between the research topics. Using data retrieved from Scopus and with D3 visualization, this paper present a technique for building such knowledge graphs which can be used to build a larger graph database of connected research keywords.

Keywords

Association Rules, Scopus, Apriori, Topics Analysis, Knowledge Graphs.

1. INTRODUCTION

Research in the world's scientific community is growing tremendously these days [1]. It became highly dynamic and fast-changing. New research areas are emerging, and interdisciplinary research is thriving [2]. Therefore, journals and research papers which covers different research subjects are increasing. Therefore, visualizing the association between research topics based on the available data can be beneficial.

Authors spend much time searching for suitable journals. Large companies are playing a vital role in managing and organizing scientific research such as Scopus [3] and Web of Science [4]. Publishers such as Elsevier, have provided tools that ease the searching process for best-fit journals depending on the article title intended for publishing [5]. Scopus also provides a tool for arranging the journals by subject and ranked by their strength in the field using quartiles categorization [6]. On the other hand, the number of available research papers is rapidly growing in interdisciplinary research. Hence, browsing journals based on their research topics they publish highlights the interdisciplinary journals emerging.

Big data will have a tremendous effect on the scholarly applications' emerging by different organizations [7]. It can be overwhelming most of the times. But when it is transformed into knowledge, it can be of a great asset in its field. This paper focuses on building knowledge graphs of associations between research topics (ASJC subjects) based on the journals and their subject topics. Scopus uses the All Subject Journal Classification (ASJC) hierarchy to denote the discipline of a journal indexed in Scopus [8]. The ASJC is a tree-like structure with no linkages established. This research paper attempts to build connections between the subjects in order to have informative information for future applications. Association rules mining using Apriori algorithm have been applied on data retrieved from Scopus API to discover binary relations between different high-level research topics.

The rest of this paper is organized as follows: background and related work explains the effect of several datamining techniques on scholarly data. The methodology proposed in this paper is explained in the following section. The dataset preparation and the experiments conducted along with the results discussion are presented in the following two sections. Before the conclusion, ideas about extension work of this research are listed for future work.

2. BACKGROUND AND RELATED WORK

Big data and data science are affecting all disciplines such as social media, e-learning, and e-commerce. As stated in [7], the outcome of big data will pave the way towards a big scholarly data platforms for the academic and scientific research. For example, recommender systems are one of the main highly researched techniques dealing with scholarly data. Research in the literature showed the use of collaborative filtering, association rules, and other novel approaches to ease the process of keyword search or browsing the web for research papers [9-12]. The availability of well-organized Application Programming Interfaces (APIs) provided by scholarly repositories made the process of retrieving and mining research data a lot easier [13]. Harvesting bibliographic metadata from Scopus and such websites, helps organizations visualize their publications, collaboration, and the impact of its research on the community [14].

In addition to several data mining techniques, association rules mining is widely used to discover interesting relations between variables in large datasets. Frequent itemsets discovered in a sequence of data can be used to generate associations rules with the use of a measure known as confidence [15]. Apriori algorithm is considered a brute force approach in which it enumerates all possible itemsets of data and then determines the most frequent ones [16]. The measure for pruning the frequent items is called the Confidence measure.

In brief, association rules mining can be explained as follows. Let X, Y be two itemsets, the confidence of a rule X→Y is the conditional probability that a transaction has the itemset Y, given that it contains the itemset X. The frequent itemsets are the frequent patterns with high confidence measure in the collection of items being studied. The higher the confidence, the more frequent that items appear together in the collection studied. Associations rules mining can benefit the research field in mining frequent patterns in large databases and help extract valuable information [17]. If applied to scholarly datasets, such method can benefit finding frequent patterns between authors, papers published, and journals. Apriori algorithm used to discover association rules is helpful in building knowledge about relations between items in any dataset [18]. The rest of the paper explains the use of Apriori to build knowledge graphs of research topics associations.

3. METHODOLOGY

In this paper, we used data harvested from Scopus via Elsevier APIs [19] to retrieve metadata about journals titles published in their dataset. This API is available upon registration and an API authorization key is provided. Data is retrieved by submitting a query string parameter using the subject criteria specified. The query returns the response in JSON or XML mark-up as specified.

The methodology for collecting the dataset and discovering patterns is explained in figure. 1. The figure details the flow of the work in steps. The main search criteria applied in the API request was the subject. An iteraative approach was used to collect data from all subjects provided by the ASJC collection. Data cleaning and preparation phase included extracting only the ISSNs and the subject IDs for each journal and convert them to a suitable format to work with using the Apriori algorithm. For each subject-based data collected, binary frequent patterns were discovered and filtered to include the rules with the highest confidence. The sets of rules are then accumulated to build the knowledge graph of subjects' associations.

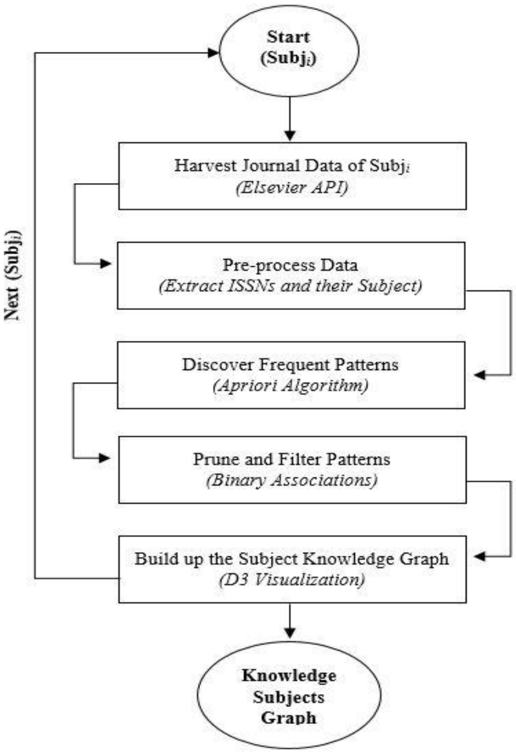
Start
(Subj$_i$)
Harvest Journal Data of Subj$_i$
(Elsevier API)
Pre-process Data
(Extract ISSNs and their Subject)
Discover Frequent Patterns
(Apriori Algorithm)
Prune and Filter Patterns
(Binary Associations)
Build up the Subject Knowledge Graph
(D3 Visualization)
Knowledge
Subjects
Graph
Next (Subj$_i$)

Figure 1: Methodology for building knowledge graph

4. DATASET PREPARATION

4.1. Dataset Collection

Regardless of the field of study, accurate data collection is a very important part of the project. It maintains the integrity of the research. Hence the Scopus data is used as a very reliable and accurate source of scholarly data. Scopus APIs have helped researchers analyze cited-by counts across academic disciplines, study relationships between authors and the publications, and many more. In order to use Scopus APIs to showcase our research, you can easily obtain a Scopus API key for free, but some restrictions are held on query views that can return detailed information about its entities (authors, articles, journals, affiliation, etc.).

For this experiment, the Serial Title API was used, and it provided an interface to search and retrieve all the serial titles indexed in Scopus. Only journal titles were retrieved for this experiment. The data set preparation included retrieving journals categorized in all 27 subjects listed in ASJC list. The subjects (research topics) are illustrated in figure 2, and it is represented in a tree like structure in the ASJC.

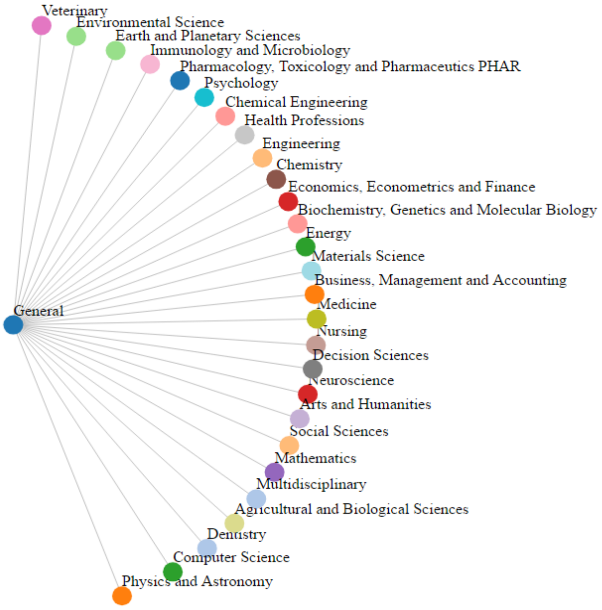

Figure 2: ASJC tree representation

Several queries were ran to retrieve journals information for each subject code in the ASJC. The data is retrieved and saved in XML or JSON format. In this experiment, the quries applied to filters for the data collection which included in the *subject*, the *type* of returned title, in order to collection only journals informations.The results were stored in JSON format and pre-processed for later usage. For each subject area (ASJC code), the data is retrieved by posting several queries in standard query view and iterated several times to retrieve all journals categorized in the subject since it has a limited number of returned items at each query response.

4.2. Dataset Pre-processing

Data preprocessing is a major step in data mining techniques because it transforms raw data into an understandable format, that can match our algorithm requirement [15].

The data retrieved, about all journals and their subject codes, have been converted into CSV format. Then, journals with a single subject in its subject list (ASJC codes) were eliminated the. For example, the field of Medicine had the largest dataset which includes 4445 journals, but 99% of the listed journals have stated publishing in one subject only, that is medicine. , The medicine journals showed less openness to interdisciplinary research in their journals. Such journals had to be eliminated from the dataset as they can affect negatively on the Apriori algorithm results. Such observation might mean that medicine specialized journals can be a bit isolated. On the other hand, the journals categorized with Engineering subject were 1460 journals, and most of them have other subjects in their categorization that belongs to sub engineering topics. All the subjects categorizing the journals were converted to their top-level subject to build a more comprehensive knowledge graph. This will have a positive effect on discovering frequent patterns for this subject area.

The same process of retrieving data, filtering and pre-processing it, and then discovering frequent patterns, have been iterated for the 27 subjects included. The resulting binary associations were then visualized in D3 graphs showing the confidence of the patterns discovered [20]. The next chapter will show detailed results and experiments conducted for two fields as a sample from the data collection and analysis process. This process was iterated to include all the subjects and the knowledge graph of association rules have been growing in each iteration.

5. EXPERIMENTATION AND DISCUSSION

The first experiment detailed in this section is showing the results of applying Apriori algorithms on the data set collected for the journals of Computer Science subject.

5.1. Computer Science

The experiments began by applying the Apriori algorithm on the data retrieved for the computer science research subject. 262 journals had been retrieved for the computer science subject. After the data pre-processing explained in the previous section, only 171 journals have been categorized in more than one subject (ASJC code) . Table 1 shows the results of the experiments conducted to discover frequent patterns with varying support values.

Table 1: Experiments details for computer science

Exp. No	Support value	All rules gen.	No of rules generated for size 2	No of rules generated for size 3
1	0.1	3	3	NA
2	0.09	5	5	NA
3	0.02	30	20	8
4	0.009	212	59	81

In this experiment, the strength of computer science association with other subject areas is discovered. Therefore, only binary associations wer retrieved using Apriori algorithm. Results are shown in Table 2 and the rules are visualized as a graph in figure 3.

Table 2: Binary rules for computer science

Lhs	Rhs	Sup.	Conf.	Count
{Social Sciences}	{Computer Science}	0.163	1	28
{Mathematics}	{Computer Science}	0.315	1	54
{Engineering}	{Computer Science}	0.497	1	85

As the results show, computer science and engineering subject areas are strongly connected based on the journals subject list provided in the dataset. Therefore, the second experiment retrieved data for the Engineering subject to start applying the methodology in an iterative approach.

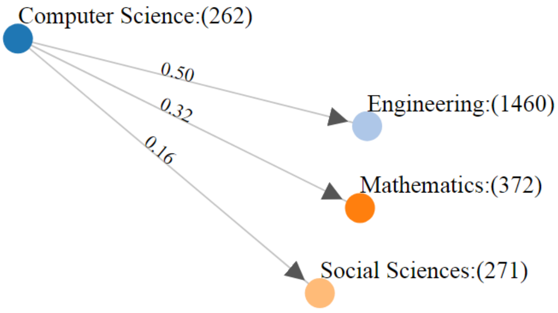

Figure 3: Computer science relationships

5.2. Engineering

The same experiments were applied on data for the journals retrieved from the engineering subject query. The serial title API returned 1460 journals for engineering subject. After cleaning the dataset and removing journals with only one subject area, only 368 journals were lift. Table 3 shows the detailed results of the experiment showing a larger number of rules generated than computer science.

Table 3: Experiments details for engineering

Exp. No	Support value	All rules gen.	No of rules generated for size 2	No of rules generated for size 3
1	0.1	10	8	2
2	0.08	12	9	3
3	0.04	16	12	4
4	0.01	51	16	27

As in the previous experiment, the rules generated at the highest level of support were used the rules of other sizes were eliminated. The results were 8 rules generated from experiment no 1, of sizes 2 detailed in Table 4.

Table 4: Binary rules for engineering

Lhs	Rhs	Sup.	Conf.	Count
{Social Sciences}	{Engineering}	0.125	1	46
{Mathematics}	{Engineering}	0.154	1	57
{Physics and Astronomy}	{Engineering}	0.173	1	64
{Materials Science}	{Engineering}	0.192	1	71
{Earth and Planetary Sciences}	{Engineering}	0.239	1	88
{Computer Science}	{Engineering}	0.255	1	94
{Environmental Science}	{Engineering}	0.260	1	81

The rules of the experiment were added to the computer science graph. It resulted in larger graph of interdisciplinary research areas shown in figure 4. The figure built on the computer science associations and added new subject areas that have high confidence rules with engineering such as environmental science.

The generated rules are visualized using directed edges that demonstrates the rules and support for each rules in both direction. For example, the association between computer science and engineering subject areas is bidirectional with higher support for this rul in the computer science journals that the engineering journals.

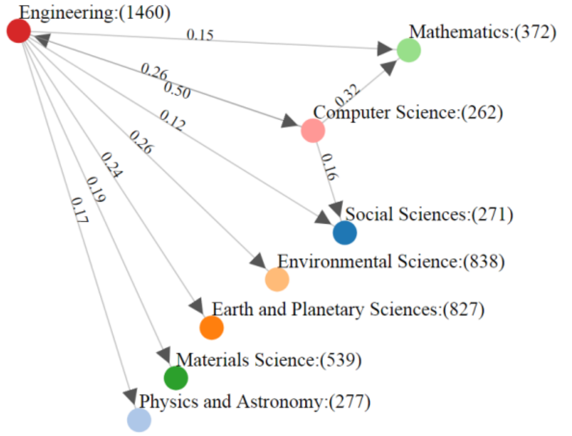

Figure 4: Engineering and computer Science relations

5.3. Environmental Science

As explained in the methodology, the method of building a complete knowledge graph is performed by iterating each new subject node and discover more associations. This experiments demonstrate the results of discovering the association for the Environmental Science which is the subject area with highest association with the engineering subject from the previous experiment. Table 5, details the experiments and number of rules generated.

Table 5: Experiments details for environmental science

Exp. No	Support value	All rules gen.	No of rules generated for size 2	No of rules generated for size 3
1	0.1	4	3	1
2	0.08	5	4	1
3	0.04	22	14	8

The binary rules generated with the highest support in the experiment above are detailed in table 6. Building on figure 4 from the previous experiment, the discovered rules added 3 more relations and one extra node that is the “Agricultural and Biological Sciences” as illustrated in figure 5.

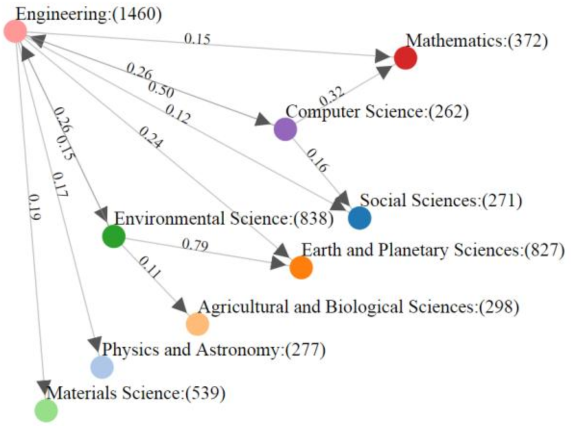

Figure 5: Adding the eniv. science subject area relations

Table 6: Binary rules for environmental Science

lhs	rhs	Sup.	Conf.	Count
Agricultural and Biological Sciences	Environmental Science	0.107	1	83
Engineering	Environmental Science	0.148	1	114
Earth and Planetary Sciences	Environmental Science	0.790	1	609

5.4. Social Sciences

Social Sciences subject area has the weakest relation with the Engineering topic. The methodology suggest iterating all generated nodes in the graphs and generate more binary association in order to discover relatinos. After experiments with the Environmental Science which resulted in adding one node to the knowledge graph generated. Social sciences is less relevant topic to the nodes already generated. Therefore, the same experiments were conducted on this subject and the results are detailed in table 7.

Table 7: experiments details for social sciences

Exp. No	Support value	All rules gen.	No of rules generated for size 2	No of rules generated for size 3
1	0.1	3	3	0
2	0.08	6	5	1
3	0.04	60	16	24

The binary rules resulted with the highest support are detaliled in table 8. As assumed, this topic introduced two new nodes to the graph and one new relation the the Enviromental science. This assures the connection between these two since it is bidirectional relation now. The new nodes introduced to the graph, as illustrated in figure 6, have strong association with the social science and no relation with any other node for now.

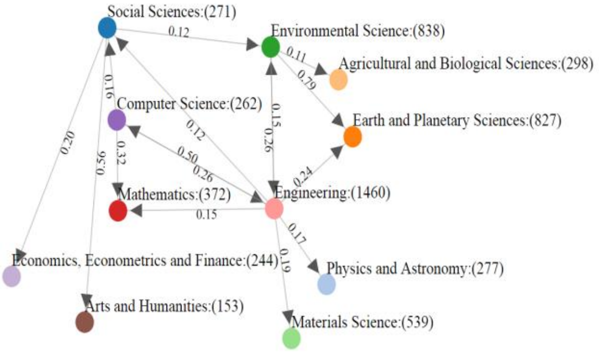
Social Sciences:(271)
0.12
Environmental Science:(838)
0.11
Agricultural and Biological Sciences:(298)
0.79
0.16
Computer Science:(262)
0.15
0.26
Earth and Planetary Sciences:(827)
0.20
0.56
0.32
0.12
0.50
0.26
0.24
Mathematics:(372)
Engineering:(1460)
0.15
0.17
Economics, Econometrics and Finance:(244)
0.19
Physics and Astronomy:(277)
Arts and Humanities:(153)
Materials Science:(539)

Figure 6: Introducing the social science and its relations

Table 8: Binary rules for social sciences

lhs	rhs	Sup.	Conf.	Count
Environmental Science	Social Sciences	0.119	1	17
Economics-Econometrics and Finance	Social Sciences	0.197	1	28
Arts and Humanities	Social Sciences	0.556	1	79

Having such initial results, the proposed algorithm showed a the ability to represent strong associations that represent the relations between different research journals in different fields. Therefore, the experiments continued in an iterative approach in order to read data for all the subjects in the ASJC and build a larger connected graph. The graph shows the relations between the 27 subjects in the ASJC based on journals publishing in multidisciplinary research topics. Figure 7, demonstrate the full relations between all the subjects in one graph . The graph is visualized using D3 tools after generating the rules for all the data collected. It is complex graph because it is representing the relation in a directed graph. Therefore the weights of support were removed to enhance the readability of the figure. For each directed relation, a weight is generated to inducate the strength of the connection as it was shown in the previous figures in this paper.

6. EXTENSION WORK

The work presented in this paper is considered the beginning of a larger project on scholarly data. Graph databases are important these days. Everything is connected in this world. If more data is represented as graphs, more applications can be developed to exploit its content. In this section more ideas are presented that can build on the work presented here.

- Building complex knowledge graphs

Using more complex data mining algorithm, such as clustering with association rules mining, interdisciplinary subjects can be easily identified, especially if enriched with semantics. ASJC subtopics in addition to keywords of articles published in the journals can enrich the process of discovering relations between high-level classification and lower level subjects in the ASJC.

- Publishing linked open data graph

Such knowledge graphs can be published in RDF format. It then can be used for annotating scholarly datasets such as researchers of organizations and their research interest. Having such a layer that can be used for annotating such data, people with similar research interests can be easily discovered. Furthermore, research papers that are interdisciplinary can benefit from such graphs especially if it was enriched with keywords to represent each subject areas. Graphs can be interactive and allow use clicks on edges to show the list of journals that are publish in different subject areas.

- An ontology representing scholarly data

The knowledge graph concluded from this relation explains a small part of a larger ontology. Each journal is categorized into one or more subject areas that have subtopics. Articles are published in journals and the articles have keywords which might be used to detect the topic or subject of the article. Each author publishes in a journal with an article. The whole scholarly data can be represented in an ontology of connected entities with relations and attributes for each entity. Such ontology can be beneficial in order to discover more hidden knowledge.

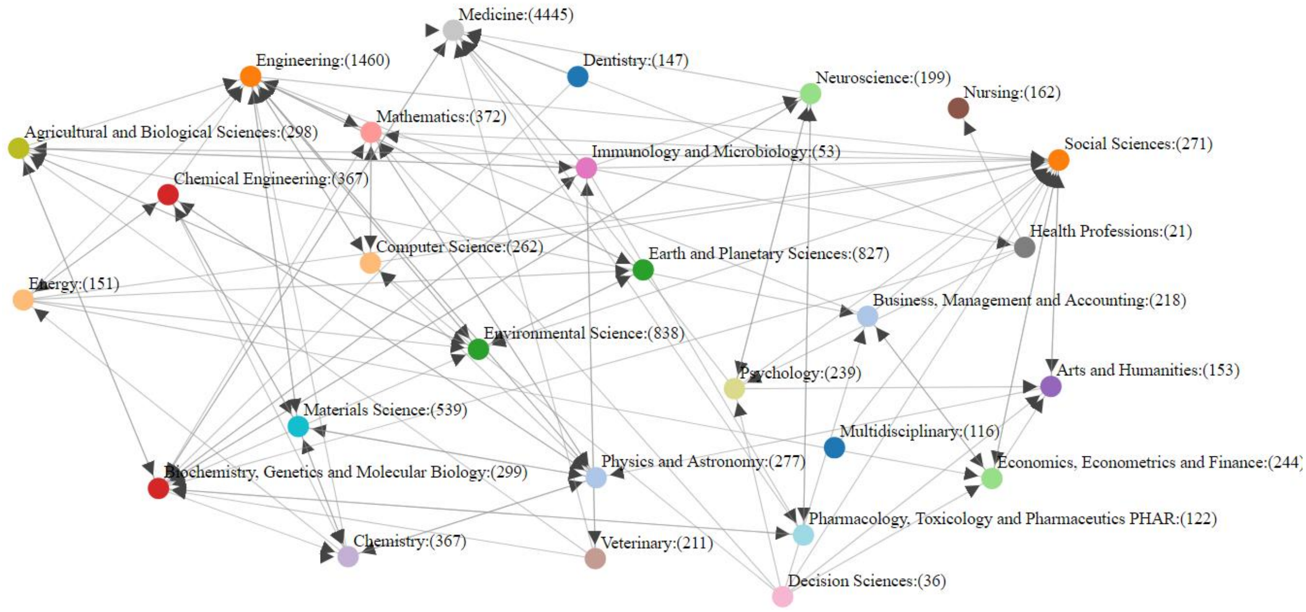
Medicine:(4445)
Engineering:(1460)
Dentistry:(147)
Neuroscience:(199)
Nursing:(162)
Agricultural and Biological Sciences:(298)
Mathematics:(372)
Immunology and Microbiology:(53)
Social Sciences:(271)
Chemical Engineering:(367)
Health Professions:(21)
Computer Science:(262)
Earth and Planetary Sciences:(827)
Energy:(151)
Business, Management and Accounting:(218)
Environmental Science:(838)
Psychology:(239)
Arts and Humanities:(153)
Materials Science:(539)
Multidisciplinary:(116)
Physics and Astronomy:(277)
Economics, Econometrics and Finance:(244)
Biochemistry, Genetics and Molecular Biology:(299)
Pharmacology, Toxicology and Pharmaceutics PHAR:(122)
Chemistry:(367)
Veterinary:(211)
Decision Sciences:(36)

Figure 7: Complete subjects' knowledge graph of association

7. CONCLUSION

The Apriori algorithm applied for frequent pattern mining was successful in discovering rules with high confidence between different subject areas classified using the ASJC data. The rules were generated based on journals that were categorized by many subject areas. The goal of building knowledge graphs that build associations between research topics was applicable using the proposed methodology presented in this paper. Though, more interactive illustrations can be beneficial for displaying the interdisciplinary journals between two areas. The limitation of the work was the small number of journals indexed in some subject areas. Integrating such a technique with all journals published in several repositories would provide the capability to produce stronger relations. This research is part of a work in progress and more ideas can build on this work as stated in the extension work section.

8. AUTHOR CONTRIBUTIONS

R. Q. A. initiated the idea, programmed, and co-authored the paper. H. S. helped developed the idea and co-authored the paper. S. S. programmed and conducted the experiments, B. A. Analyzed and visualized the results. All authors had approved the final version.

9. REFERENCES

[1] A. A. Chadegani, H. Salehi, M. M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, & Nader, and A. Ebrahim, "A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases," Asian Soc. Sci., vol. 9, no. 5, 2013.

[2] K. K. Mane and K. Börner, "Mapping topics and topic bursts in PNAS.," Proc. Natl. Acad. Sci. U. S. A., vol. 101 Suppl 1, no. suppl 1, pp. 5287–90, Apr. 2004.

[3] "Scopus preview - Scopus - Welcome to Scopus." [Online]. Available: https://www.scopus.com/home.uri. [Accessed: 20-Sep-2019].

[4] "Web of Science - Please Sign In to Access Web of Science." [Online]. Available: http://login.webofknowledge.com/. [Accessed: 25-Sep-2019].

[5] "Find journals | Elsevier® JournalFinder." [Online]. Available: https://journalfinder.elsevier.com/. [Accessed: 25-Sep-2019].

[6] "SJR : Scientific Journal Rankings." [Online]. Available: https://www.scimagojr.com/journalrank.php. [Accessed: 25-Sep-2019].

[7] S. Khan, X. Liu, K. A. Shakil, and M. Alam, "A survey on scholarly data: From big data perspective," Inf. Process. Manag., vol. 53, no. 4, pp. 923–944, Jul. 2017.

[8] "What is the complete list of Scopus Subject Areas and All Science Journal Classification Codes (ASJC)? - Scopus: Access and use Support Center."[Online].Available:https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/related/1/session/L2F2LzEvdGltZS8xNTY5NDM2MjAyL2dlbi8xNTY5NDM2MjAyL3NpZC9mVUFGUE9jRlNtZWRRTG9HZUVFZW9tU2xRek5nWUVueTh5dkJ0VVYzY21xdnB6dFA3UEx6JTdFU2JZdG85VjVTMmlZZjJScEtlQ1JqUG1f. [Accessed: 25-Sep-2019].

[9] F. Xia, H. Liu, I. Lee, and L. Cao, "Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences," IEEE Trans. Big Data, vol. 2, no. 2, pp. 101–112, Jun. 2016.

[10] J. Lee, K. Lee, and J. G. Kim, "Personalized Academic Research Paper Recommendation System,", Information Retrieval, Apr. 2013 [online].

[11] A. Naak, H. Hage, and E. Aïmeur, "A Multi-criteria Collaborative Filtering Approach for Research Paper Recommendation in Papyres," in MCETECH 2009: E-Technologies: Innovation in an Open World, 2009, pp. 25–39.

[12] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, 2011, p. 448.

[13] C. C. Austin, S. Brown, N. Fong, C. Humphrey, A. Leahey, and P. Webster, "Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements," 2015.

[14] W. H. Mischo, M. C. Schlembach, and E. Cabada, "Visualizing the Scholarly Impact of Medical Education Researchers," Qual. Quant. Methods Libr., vol. 8, no. 2, pp. 169–178, Sep. 2019.

[15] C. C. Aggarwal, An Introduction to Data Mining. Cham: Springer International Publishing, 2015.

[16] M. J. Zaki and W. Meira, Data mining and analysis : fundamental concepts and algorithms. 2014.

[17] M. Abdel-Basset, M. Mohamed, F. Smarandache, and V. Chang, "Neutrosophic Association Rule Mining Algorithm for Big Data Analysis," Symmetry (Basel)., vol. 10, no. 4, p. 106, Apr. 2018.

[18] E. Petrova, P. Pauwels, K. Svidt, and R. L. Jensen, "In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data," in Advances in Informatics and Computing in Civil and Construction Engineering, Cham: Springer International Publishing, 2019, pp. 19–26.

[19] Elsevier Developers, "Serial Title API." [Online]. Available: https://dev.elsevier.com/documentation/SerialTitleAPI.wadl. [Accessed: 25-Sep-2019].

[20] Fan Bao and Jia Chen, "Visual framework for big data in d3.js," in 2014 IEEE Workshop on Electronics, Computer and Applications, 2014, pp. 47–50.