Finding Association Rules in Linked Data, a Centralization Approach

Reza Ramezani
Electrical & Computer Engineering
Isfahan University of Technology, Iran
r.ramezani@ec.iut.ac.ir

Mohammad Saraee
School of Computing, Science and Engineering
University of Salford, Manchester, UK
m.saraee@salford.ac.uk

Mohammad Ali Nematbakhsh
Department of Computer Engineering
University of Isfahan, Iran
nematbakhsh@eng.ui.ac.ir

Abstract- ***Linked Data is used in the Web to create typed links between data from different sources. Connecting diffused data by using these links provides new data which could be employed in different applications. Association Rules Mining (ARM) is a data mining technique which aims to find interesting patterns and rules from a large set of data. In this paper, the problem of applying association rules mining using Linked Data in centralization approach has been addressed -i.e. arranging collected data from different data sources into a single dataset and then apply ARM on the generated dataset. Firstly, a number of challenges in collecting data from Linked Data have been presented, followed by applying the ARM using the dataset of connected data sources. Preliminary experiments have been performed on this semantic data showing promising results and proving the efficiency, robust, and useful of the used approach.***

Keywords: Data Mining, Association Rules Mining, Linked Data Mining, Frequent Itemset Mining, Linked Data Query.

I. Introduction

Linked Data is a practical technique used to achieve Semantic Web goals [2]. Linked Data can be defined as "*a set of best practices for publishing and connecting structured data on the Web*" [2]. Linked Data has goal to connect semantic web datasets together and hence many natural features of semantic web data, exists also in Linked Data. Extending the scope of data mining research from traditional data to semantic web data with its Linked Data could help to discover and mine richer and more useful knowledge [3, 4].

ARM, one of the main data mining techniques, tries to find frequent itemsets and based on these frequent itemsets, generates interesting association rules (ARs).

The problem of mining ARs from semantic web data is faced to several challenges, such as: *heteronomous data structure, no exact definition of transactions in semantic web data, relationships between entities and role of end user in mining process*. In [1] an algorithm, named SWApriori, has been proposed which has considered the above challenges and without the end user involvement, mines ARs directly from a single semantic web dataset. So we assumed that the mentioned problem has been solved.

In this paper, the problems and challenges of collecting desired data, from different Linked Data sources, and the ARM of it has been investigated. The presented approach collects data from various data sources and connects them so that all data appear as a single and central dataset and then uses existing methods [1, 5] to mine ARs from the generated dataset.

The rest of the paper is organized as follows: section 2 introduces a number of related work in this area. Section 3 expresses the problem of mining ARs from semantic web data and introduces two existing methods. Section 4 addresses some of the challenges of mining ARs from Linked Data, describes a number of methods for extracting Linked Data, and finally describes the selected solutions to solve these challenges. Section 5 introduces the datasets used with a list of numerical statistics. Section 6 gives the experimental results and finally Section 7 presents our conclusion and recommendations for future work.

II. Related Work

The ARM was first introduced in [6, 7]. So far many ARM algorithms have been proposed which could work with traditional datasets [8-10]. These algorithms can be classified into two main categories: Apriori based [11, 12] and FP-Tree based [13-15].

The contents of Linked Data datasets are convertible to a directed graph with labeled edges. Other related work are the use of frequent sub-graph and frequent sub-tree techniques for pattern discovery from graph structured data [13, 14, 16, 17]. These methods are not appropriate for mining ARs from Linked Data. Because these methods are based on transactions and as was mentioned in [1] in semantic web data and along with it in Linked Data there is no exact definition of transactions and also after converting datasets contents to a single graph, each vertex of the graph, independent of its incoming link, will not be replicated in the entire graph more than once. On the other hand graph vertices are unique and thus discovering sub-graphs redundancy isn't possible.

All above work deal with traditional data. In [5] an algorithm has been introduced that mines ARs from semantic web data via mining patterns which the end user provides. The mining patterns are based on SPARQL. This algorithm is user centric and mines ARs by users' help in a semi-supervised manner. In contrary, the presented algorithm in [1] mines ARs from a single and centralized semantic web dataset without the end users' involvement and in unsupervised manner.

One of the recent works on Linked Data mining is LiDDM [18]. LiDDM is a software which is able to apply data mining techniques (clustering, classification and association rules) on Linked Data [2]. The working process behind LiDDM is as follows: firstly, the software acquires required data from LOD datasets by using user-defined queries. This followed by combining the returned results by using common columns or by appending results and then converts them to traditional data in a tabular format. At the next level, a number of pre-processing steps will be applied on these data. Finally, traditional data mining algorithms will be applied. The different between the proposed technique and LiDDM are in the method used to acquire and combine data and mining approach of ARs from connected data. LiDDM uses traditional data mining algorithms which data must be converted to traditional data format (tabular format). While in the proposed method, data remain in its semantic web format, i.e. triples (subject, predicate, object).

RapidMiner semweb plugin [19] is a similar approach to LiDDM which applies data mining techniques on semantic

978-1-4673-5634-3/13/$31.00 ©2013 IEEE

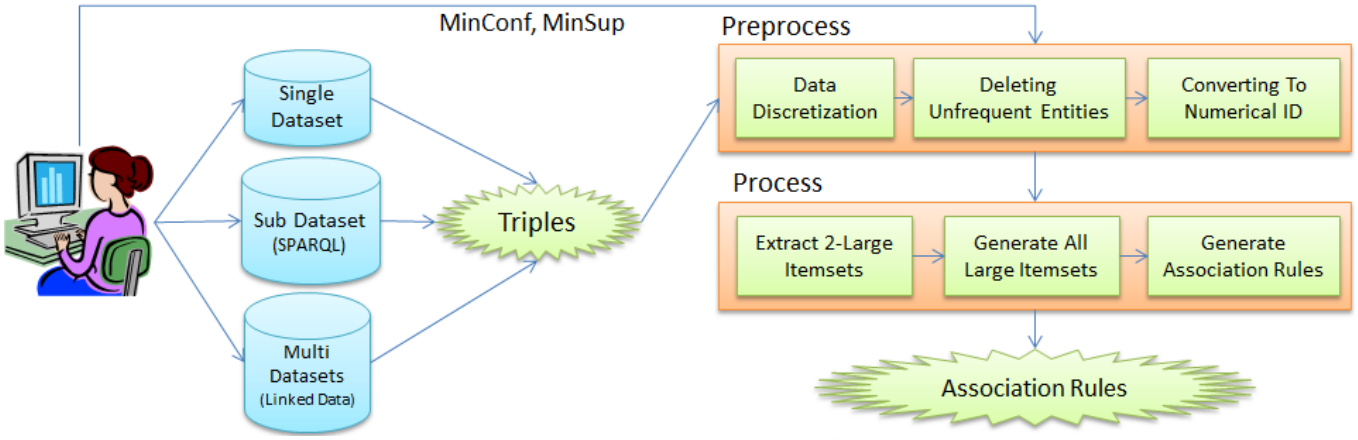

Figure 1 – SWApriori Mining Process Workflow [1]

web data or Linked Data. Similar to LiDDM, in RapidMiner semweb plugin, the end user has to define a suitable SPARQL query for retrieving desired data from Linked Data datasets and converts data to tabular feature format. This conversion is not appropriate for heterogeneous datasets because it causes a loss in the data or the appearance of abundance null values.

In RDF structure, each statement names triple and is identified with three parts: *subject*, *predicate* and *object*. In order to generate transactions from semantic web data, it's possible to use one of these three values to group transactions (TID) and to use one of the other remaining values as transaction items and then mines ARs from these transactions. Six different combinations of these values along with their usage are shown in Table 1 [20]. For example, grouping triples by predicate and using objects for generating transactions has usage in clustering.

TABLE 1 - COMBINATIONS OF TRIPLE PARTS [20]

	Context	Target	Use Case
1	Subject	Predicate	Schema discovery
2	Subject	Object	Basket analysis
3	Predicate	Subject	Clustering
4	Predicate	Object	Range discovery
5	Object	Subject	Topical clustering
6	Object	Predicate	Schema matching

SPARQL-ML [21] is another approach to mine semantic web data that provides special statement as an extension to SPARQL query language to create and learn a model for specific concept of retrieved data. It applies classification and regression techniques on data, but other data mining techniques such as clustering and association rule aren't covered by this approach.

III. MINING ASSOCIATION RULES FROM SEMANTIC WEB DATA

This paper concentrates on the problem of mining ARs from Linked Data. First of all, the problem of mining ARs from semantic web data will be investigated since the proposed approach is based on collecting desired and related data from different datasets by considering existing challenges, putting collected data in a central dataset and then using one of the existing methods to mine ARs from this single dataset. We call this proposed method as *centralization approach*.

Two important work about mining ARs from semantic web data are [5] and [1]. The first one [5] is using a provided end user's query pattern to extract and generate transactions from a single semantic web dataset and then mine ARs from these transactions by applying traditional ARM algorithms. The second one [1] is proposed a new algorithm named **SWApriori** that receives a semantic web dataset in triple format and then launches to mine ARs from this dataset. This algorithm mines ARs without the end user involvement and without any transactions extraction from semantic web data or conversion of semantic web data to traditional data. In fact, SWApriori receives a dataset, which contains triples, as input and automatically mines ARs from this dataset without end user involvement.

In this paper, we assume that after collecting data from difference data sources and creating a new dataset, ARs would be mined by using SWApriori [1]. Figure 1 shows the workflow of **SWApriori** mining process. Hence the focus of this work is more on the problem of *"how to collect desired and related data from Linked Data datasets and how to connect them"* instead of the problem of *"how to generate transactions from Linked Data"* or *"how to mine ARs from Linked Data"*.

IV. MINING ASSOCIATION RULES FROM LINKED DATA

A) Challenges

There are several challenges in extracting data and also in mining ARs from Linked Data that needs to be considered. These challenges requires the mining ARs from Linked Data to be indirect, this is done by manually collect desired data from different datasets in Linked Data space, convert them to a single dataset and mine association rule from this new centralized dataset by using the proposed algorithm in [1]. The challenges are as follows:

- ✓ ***Generalized association rules:*** Linked Data datasets are categorized into two groups. Those which belong to special domain and contain information only about that domain such as Factbook that only has information about countries. Such datasets called *specialized* datasets. And those datasets that don't belong to specific domain and have data about various domains. Such as DBPedia which contains data about many concepts, such as people, books, music, geography, medical and etc. Such datasets called *generalized datasets*. The quality of the generated ARs depends on dataset contents. If dataset's contents are generalized and don't belong to specific domain, the generated ARs would be generalized too (i.e. the parts of rules are from different domains) and thus are approximately unmeaning. Moreover other meaningful rules have low frequency (Support) rather than number of entities. Thus to mine these rules, the *MinSup* value has to be decreased adequately. As a result, the number of generated rules would be increased respectively. *While deal with generalized datasets, we have to focus on*

special domain and only use a subset of generalized datasets that is related to a desired and specific domain.

- ✓ ***Large amount of existing Linked Data:*** Usually in a semantic web dataset (whether generalized or specialized) there is a large amount of data (triples). Nowadays, there are many datasets in LOD project that connect semantic web datasets together and increase the amount of all triples respectively. And that's why using all LOD datasets isn't reasonable decision. A challenge is to decide what datasets and what parts of these datasets must be selected. Another challenge that may be raised is selecting the start point of data collecting and the criteria of connecting data sources. *These challenges are related to the connecting dataset mechanism.*
- ✓ ***Different ontologies:*** Usually each dataset uses one or more ontologies to describe its data. Hence different datasets usually have different ontologies. As a result, it is possible to have identical data with different names or different data with identical name. This problem needs to be considered while connecting or merging datasets. *One solution is to use ontology mapping concepts [22-24].*
- ✓ ***Duplicated data:*** This problem occurs in two forms. The first one is when there are two identical subjects and predicates in different datasets with different objects value. For example a dataset may show Iran population as 75,000,000 people and another may show 76,000,000 people. The second one is when there are multiple object values for a predicate of a subject. *Only one of the duplicated values (the valid value) must be selected in the mining process.*
- ✓ ***Lack of access to a data source data:*** As the first step of data mining, desired data have to be extracted from data sources. Unfortunately some Linked Data datasets don't provide any mechanism to directly access, traverse and extract desired data. Many data sources allow to download entire dataset or to retrieve desired data by using SPARQL. In contrast some datasets neither allows downloading entire dataset and neither supports SPARQL. In the worst case, a dataset only provides data in HTML based format. *In this situation, it's possible to use data, but these data have to be converted to a suitable format. Also in this case the ontologies would be lost.*

B) Data extraction methods

In this part, a number of methods for extracting desired data from Linked Data will be discussed. It's valuable to mention that in this paper the concept of data extraction is different to the concept of data query. The problem of querying Linked Data has been addressed in [25]. Four important techniques for querying Linked Data are: entity-centric (SWSE/Visinav [26], Sindice/Sigma [27]), structure indexes (Semplore [28], Dong and Halevy [29]), question answering systems (PowerAqua [30], Freya [31]) and natural language search (Treo [32], Treo T-Space [33]). The difference between querying data and extracting data is that the goal of data query is to provide information as a result of applying a special request/query while data extraction goal is to provide data as the data source for data mining. A number of methods of data extraction are described as follows:

- ✓ ***Connecting complete datasets:*** The simplest and the most straightforward way for collecting data from different sources is to append the contents of a dataset to the end of another dataset. This means that the triples of a dataset are placed after another dataset's triples. This approach requires the availability of entire (part of) dataset by downloading (querying) desired dataset.
- ✓ ***Data extraction by SPARQL:*** Another method is to use SPARQL as Linked Data query language. In data sources that provide SPARQL endpoint, it's possible to extract desired data by using suitable SPARQL commands. Querying Linked Data can be done in three ways. The first way is to query different data sources independently and then merge or append acquired data. The second method is to follow-up queries over other data sources based on the results from previous queries and substituting placeholders in query templates. The third method is to use an existing SPARQL endpoint that provides access to a set of copies of relevant data sources such as, SPARQL endpoint by OpenLink SW which has access to the majority of data sources from the LOD cloud at http://lod.openlinksw.com/sparql [34].
- ✓ ***Automatic HTML traverser:*** Another possible but not suitable method is to traverse automatically HTML web pages that contain desired data and then convert them to appropriate format. This method is applicable but not appropriate since the traverser source code has to be tuned for each dataset provider (HTML Web page) as well as the elimination of ontology concepts from extracted data. This method is useful when a data source doesn't provide any data access method.
- ✓ ***Automatic data sources traverser:*** The goal of data extraction is to provide desired data to apply data mining techniques on them. In Linked Data, in order to collect data, the best method is to determine desired domain and may be some criteria of appropriate data and then a data extractor will automatically traverse available datasets and extract desired data from them. This traverser can select datasets that match the determined criteria, extract related data and finally connect or merge all extracted data from different datasets to each other. As another approach, this data extractor is selecting a dataset as the start point of data extraction followed by collecting related data from this dataset and then traverse suitable other datasets by using existing knowledge in the so far extracted data. For example by using *owl:sameAs* or *owl:seeAlso* predicates [35]. Regardless to the type of this method, the automatic data sources traversers need to be able to deal with different datasets that have different ontologies and different structures. In these traversers, the problem of inaccessibility of a dataset's contents must be solved.

C) Selected Solution

In the two previous parts, some challenges in mining ARs from Linked Data and some methods for extracting desired data were discussed. In this paper, in order to mine ARs from Linked Data, desired data has been collected from two datasets (DBPedia and Factbook) and then converts them to a singled and central dataset and finally mines ARs by using the proposed algorithm in [1].

Next sections show the datasets used and acquired results. This part describes the used methodology of collecting desired data.

- ✓ ***Selecting Domain:*** As mentioned earlier, generated ARs quality depends on the input dataset quality and how much the provided data is being specialized in a domain, the generated rules are more specific and more useful

respectively. Also because generalized data are diffused, the *MinSup* value needs to be low so that these generalized rules appear in the result. In this paper we selected ***Country*** as data domain.

✓ ***Selecting Datasets:*** Next step is to select suitable datasets that have appropriate data about countries (the selected domain). As mentioned in section B), there are two approaches to automatically select suitable datasets. Our strategy for selecting datasets is to consider a dataset as the start point and extract desired data from it, afterward select another datasets based on the extracted data and finally traverse these new selected datasets. This operation will be continued until a special criterion is being satisfied (for example traverse at most 5 datasets). The selected dataset for start point is very important. In addition to have many links to other datasets, the start point dataset must have suitable data about selected domain. ***DBPedia***[1] has been selected as start point dataset since it has many links to other datasets.

✓ ***Connecting Datasets:*** As mentioned before, there are two ways for connecting related data of different datasets. The first way is to acquire suitable data independently and then append them to each other or connect them by using common attribute [18]. The second way is to collect data from a dataset and then traverse another datasets by using objects from triples of collected data. The latter method requires an explicit relation among datasets- i.e. object of a triple refers to an entity (a subject) that exists in another dataset. These explicit relations are grouped into two relations' groups: *more information* relations and *equivalent* relations. More information relations are those relations that state one attribute of an entity which has been defined and exist in another dataset. For example suppose entity *PersianGulf* has been defined in a dataset (DS1) and *Iran* has been defined in another dataset (DS2). Suppose this triple exists in DS1:

DS1:PersianGulf MyOnto:IsMainPartOf DS2:Iranin

In this example, *MyOnto:IsMainPartOf* is a more information relation. Equivalent relations are those relations that state two entities in two different datasets are equivalent. The most common equivalent relation is *owl:sameAs*. It's clear that using more information relations don't help to collect suitable data about a domain since these relations only show external attributes causing irrelevant entities in the collection process. This is why only equivalent relations have been used to collect suitable and relevant data after collecting primary data from start point dataset.

✓ ***Ontology Mapping:*** As mentioned before, due to existence of multiple data sources, ontology mapping must be applied on data so data become coherent. There are many approaches do ontology mapping [22-24]. In this paper, ontology mapping has been done manually so that data become suitable for ARM by integrating subjects and predicates- i.e. detecting identical subjects or predicates with different names, and also to scale numerical objects.

✓ ***Duplicated Data:*** This problem was addressed earlier and the solution is to select the best and the most valid data and eliminate other duplicates. Verifying duplicated objects of different data sources shows that data of those data sources that belong to special domain (specialized datasets) are nearer to real world facts rather than generalized data sources. In this paper, this issue has been solved by selecting data from specialized dataset and eliminating other duplicated data.

✓ ***Last step:*** After collecting desired data and dealing with ontology mapping and duplicated data, the data have to be placed in a single and central dataset with a unified ontology. Afterward, the proposed algorithm in [1] mines ARs from the generated dataset.

V. DATASETS

This section describes the used datasets while the results of mining ARs from these datasets will be described in the next section.

As mentioned earlier, the selected domain for mining ARs is country. Two popular datasets have been used: DBPedia and Factbook, where the first one is a generalized dataset that has data about many domains and links to many other datasets while the second one is a specialized dataset about countries. DBPedia dataset is selected as the start point dataset.

A) DBPedia

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web. The English version of the DBpedia knowledge base currently describes 3.77 million things, out of which 2.35 million are classified in a consistent ontology, including 764,000 persons, 573,000 places (including 387,000 populated places), 333,000 creative works (including 112,000 music albums, 72,000 films and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases [36].

The process of collecting data started from this dataset and another datasets traversed by using *owl:sameAs* predicates of DBPedia and generating appropriate SPARQL command.

DBPedia is a generalized dataset and only a subset of this dataset has to be used. Hence as the first step, the triples that are relevant to countries were extracted – i.e. all triples of all countries. This extraction was done by SPARQL command as shown in Figure 2:

```
SELECT * {
  ?Subject rdf:type <http://dbpedia.org/ontology/Country> .
  ?Subject ?Predicate ?Object
}
ORDER BY ?Subject
```

Figure 2 – SPARQL command for retrieving countries from DBPedia

This SPARQL command extracts all triples of all countries. After eliminating unusable data and infrequent triples and also data discretization, a dataset with the following numerical statistics was extracted:

- ✓ *Total Triples: 18,480*
- ✓ *Total Distinct Subjects: 255*
- ✓ *Total Distinct Predicates: 204*
- ✓ *Total Distinct Objects: 1,169*
- ✓ *Average Predicates for each Subject: 72.47*
- ✓ *Average Predicates for each Object: 15.8*

B) Factbook

As mentioned earlier, the selected strategy to traverse datasets is to extract triples from start point dataset, use those triples that have *owl:sameAs* predicate (equivalent relation) and then traverse suitable datasets and collect relevant data by using object values. Each extracted country from DBPedia has in average 18 triples with *owl:sameAs* predicate that from

[1] http://dbpedia.org

these triples, about 12 triples refer to the same DBPedia but with different language and about 6 triples refer to data sources such as: Eurostat Linked Data[2], OpenEI[3], Freebase[4], OpenCyc[5], LinkedGeoData[6] and Factbook. By verifying these datasets' data, it became clear that Factbook's data are suitable for ARM and easy to extract, hence only those extracted triples from DBPedia that have *owl:sameAs* as predicate and Factbook as object destination were used to traverse Factbook dataset. The algorithm generates suitable SPARQL commands by traversing these triples and extracting relevant data from Factbook by using these commands. Finally after eliminating unusable data and infrequent triples and also data discretization, a dataset with the following numerical statistics was extracted from Factbook:

- ✓ *Total Triples: 24,427*
- ✓ *Total Distinct Subjects: 252*
- ✓ *Total Distinct Predicates: 131*
- ✓ *Total Distinct Objects: 856*
- ✓ *Average Predicates for each Subject: 96.93*
- ✓ *Average Predicates for each Object: 28.53*

C) Final and Centralized Dataset

After collecting data from different data sources, they have to be placed in a single and central dataset so the proposed algorithm in [1] generates ARs from this new centralized dataset. Before merging acquired data into a single dataset, ontology mapping and duplicated data elimination are applied on data. Finally a dataset with the following numerical statistics was extracted from DBPedia and Factbook:

- ✓ *Total Triples: 38,736*
- ✓ *Total Distinct Subjects: 255*
- ✓ *Total Distinct Predicates: 298*
- ✓ *Total Distinct Objects: 1,897*
- ✓ *Average Predicates for each Subject: 151.9*
- ✓ *Average Predicates for each Object: 20.41*

This dataset is feed to the proposed algorithm in [1] and some ARs are generated from this dataset.

D) Experimental Results

As mentioned earlier, the used strategy for mining ARs from Linked Data is to collect suitable data from Linked Data space by considering Linked Data query challenges, applying ontology mapping and the duplicated data elimination techniques on the collected data and then placing these collected data into a single and central dataset with a unified ontology. Finally, by using the **SWApriori** algorithm [1] ARs are generated from the dataset. In fact, SWApriori algorithm indirectly mines ARs from Linked Data, if suitable dataset is provided.

The main goal of this work is to show that Linked Data help in improving the quality of data mining results. Results show that connecting datasets causes to generate combinatorial rules- i.e. rules that their parts (antecedent and consequent) belong to different datasets.

E) Dataset

Data mining results is dependent on the input data quality. By using Linked Data, it's possible to provide more and better data for data mining and hence improve data mining results. SWApriori algorithm [1] receives a dataset as input and mines ARs in an unsupervised manner. In Section 5 a number of challenges of Linked Data query was addressed. By considering these challenges, suitable data have been collected from two data sources: DBPedia and Factbook. DBPedia is a generalized data source which contains information about many domains and has more than 1.89 billion triples[7]. Only a subset of this data source that relates to countries was used. Factbook is a specialized dataset about countries that only those countries that were referenced by DBPedia by *owl:sameAs* predicate were used. After eliminating unusable data and infrequent triples and also data discretization, DBPedia triples count reduced to 18,480 and Factbook triples count reduced to 24,427. This acquired data were merged into a single and central dataset with unified ontology so SWApriori can mines ARs from this new created dataset. After applying ontology mapping and eliminating the duplicated data, new created dataset triples count reduced to 38,736 – i.e. ontology mapping and duplicated data elimination caused to delete 4,171 triples.

F) Results

This part describes the acquired results. Table 2 shows acquired results of mining ARs from the created and centralized dataset. This dataset contains data about countries by connecting desired data from DBPedia and Factbook.

For each *Minimum Support Value* (MinSup), Table 2 shows the number of large itemsets, the number of generated rules, the confidence average of generated rules, the number of extracted rules from DBPedia, the number of extracted rules from Factbook, and the number of rules that have been generated by connecting DBPedia and Factbook together.

As shown in the results, number of extracted rules from DBPedia is less than Factbook's rules and this is due to the DBPedia data (generalized). Nevertheless, connecting datasets will generate new rules that are not in a single dataset. In fact, generalized data of DBPedia for ARM has been mitigated by connecting dataset and acquiring new data.

These results show that connecting semantic web datasets and using Linked Data could contribute to data mining process in improving the acquired result in terms of quality and quantity.

In the presented results in Table 2, the *MinConf* value is 0.7 and the *MinSup* values' range is between 0.54 and 0.8.

VI. Conclusions and Future Work

"Linked Data is simply about using the Web to create typed links between data from different sources" [2]. Using these links and connecting datasets, provides new connected data that can be used in data mining. In this paper, we addressed the link data query challenges and the problem of mining ARs from Linked Data. Then, by considering these challenges we launched to extract and connect desired data and put them in a single and central dataset. Finally, by using the proposed algorithm in [1] ARs were mined from this new created dataset. Results show that Linked Data can help to improve data mining quality and quantity.

For future work, mining ARs from Linked Data by a distributed approach is suggested- i.e. proposing an algorithm that traverses desired datasets automatically and mines ARs online and distributed instead of collecting data to a central point and then performing data mining.

[2] http://eurostat.linked-statistics.org
[3] http://openei.org
[4] http://rdf.freebase.com/
[5] http://sw.cyc.com/
[6] linkedgeodata.org

[7] This value was checked while writing paper

TABLE 2 – RESULTS OF MINING ASSOCATION RULES FROM LINKED DATA (DBPEDIA AND FACTBOOK)

Support	Large Itemsets	Rules Count	Average Confidence	DBPedia Rules	Factbook Rules	Combination Rules
0.54	228943	1315999	0.960	193 (0.01%)	1195605 (90.85%)	120201 (9.13%)
0.56	86154	456429	0.959	144 (0.03%)	419458 (91.89%)	36827 (8.06%)
0.58	32963	159428	0.958	137 (0.08%)	148008 (92.83%)	11283 (7.07%)
0.6	12966	57352	0.957	77 (0.13%)	53619 (93.49%)	3656 (6.37%)
0.62	4387	17207	0.955	39 (0.22%)	16227 (94.30%)	941 (5.46%)
0.64	1779	6300	0.954	27 (0.42%)	5872 (93.20%)	401 (6.36%)
0.66	776	2491	0.953	20 (0.80%)	2268 (91.04%)	203 (8.14%)
0.68	293	832	0.947	16 (1.92%)	744 (89.42%)	72 (8.65%)
0.7	128	331	0.946	6 (1.81%)	294 (88.82%)	31 (9.36%)
0.72	52	118	0.938	2 (1.69%)	97 (82.20%)	19 (16.10)
0.74	24	52	0.932	2 (3.84%)	43 (82.69%)	7 (13.46%)
0.76	9	18	0.928	2 (11.11%)	14 (77.77%)	2 (11.11)
0.78	4	8	0.923	0 (0%)	8 (100%)	0 (0%)
0.8	1	2	0.926	0 (0%)	2 (100%)	0 (0%)

REFERENCES

[1] R. Ramezani, "Finding Association Rules in Linked Data," M.Sc. Thesis, Computer & Electerical Engineering Department, Isfahan University of Technology, Isfahan, Iran, 2012, The proposed algorithm for ARM from semantic web data (SWApriori) will apear in a paper titled "A New approach to mining Association Rules from Semantic Web data". This paper is under publication.

[2] T. H. C.Bizer, T.Berners-Lee, "Linked Data - the story so far," *International Journal on Semantic Web and Information Systems,* pp. 1-22, 2009.

[3] A. H. G.Stumme, B.Berendt, "Semantic web mining: state of the art and future directions," *Web Semantics: Science, Services and Agents on the World Wide Web,* pp. 124-143, 2006.

[4] N. G.-P. J.M.Benitez, F.Herrera, "Special issue on "New Trends in Data Mining" NTDM," *Knowledge-Based Systems,* pp. 1-2, 2012.

[5] R. B. V.Nebot, "Finding association rules in semantic web data," *Knowledge-Based Systems,* pp. 51-62, 2012.

[6] T. I. R.Agrawal, A.N.Swami, "Mining association rules between sets of items in large databases," presented at the SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data 1993.

[7] W. F. Gregory Piateski, *Knowledge Discovery in Databases* MIT Press Cambridge, MA, USA, 1991.

[8] U. G. Hipp Jochen, and Gholamreza Nakhaeizadeh, "Algorithms for association rule mining—a general survey and comparison," presented at the ACM SIGKDD Explorations Newsletter, 2000.

[9] C. Zhang, and Shichao Zhang, *Association rule mining: models and algorithms*: Springer-Verlag, 2002.

[10] C. Hidber, *Online association rule mining* vol. 28: ACM, 1999.

[11] R. S. R.Agrawal, "Fast algorithms for mining association rules," presented at the In Proceeding of 20th international conference in large databases, 1994.

[12] K. Z. X.Liu, W.Pedrycz, "An improved association rules mining method," *Expert Systems,* pp. 1362-1374, 2012.

[13] G. K. M.Kuramochi, "Frequent Subgraph Discovery," presented at the Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 2001.

[14] S. N. Y.Chi, R.R. Muntz, J.N.Kok, "Frequent Subtree Mining - An Overview," *Fundamenta Informations,* vol. 66, pp. 161 - 198, 2005.

[15] A. R. Islam, and Tae-Sun Chung, "An Improved Frequent Pattern Tree Based Association Rule Mining Technique," presented at the Information Science and Applications (ICISA), International Conference on, 2011.

[16] V. V. Rao, and E. Rambabu, "Association rule mining using FPTree as directed acyclic graph," presented at the Advances in Engineering, Science and Management (ICAESM), International Conference on, 2012.

[17] V. T. Vivek Tiwari, S.Gupta, R.Tiwari, "Association Rule Mining: A Graph Based Approach for Mining Frequent Itemsets," presented at the Networking and Information Technology (ICNIT), 2010 International Conference on, 2010.

[18] R. I. V.Narasimha, O.P.Vyas, "LiDDM: A Data Mining System for Linked Data," presented at the Proceedings of the LDOW2011, Hyderabad, India, 2009.

[19] M. A. Khan, Gunnar Aastrand Grimnes, and Andreas Dengel, "Two pre-processing operators for improved learning from semantic web data," presented at the In First RapidMiner Community Meeting And Conference (RCOMM), 2010.

[20] F. N. Ziawasch Abedjan, "Context and Target Configurations for Mining RDF Data," presented at the SMER '11 Proceedings of the 1st international workshop on Search and mining entity-relationship data 2011.

[21] A. B. Christoph Kiefer, André Locher, "Adding data mining support to SPARQL via statistical relational learning," in *ESWC'08 Proceedings of the 5th European semantic web conference on The semantic web: research and applications methods*, 2008, pp. 478-492

[22] L. Huang, Guoxiong Hu, and Xinghe Yang, "Review of ontology mapping," presented at the Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on. IEEE, 2012.

[23] I.-Y. S. Namyoun Choi, Hyoil Han, "A Survey on Ontology Mapping," presented at the ACM SIGMOD Record, 2006.

[24] M. S. Yannis Kalfoglou, "Ontology mapping: the state of the art," *The Knowledge Engineering Review,* vol. 18, pp. 1 - 31, 2003.

[25] A. C. Freitas, E. ; Oliveira, J.G. ; O'Riain, S, "Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends," *Internet Computing, IEEE,* vol. 16, pp. 24-33, 2012.

[26] A. Hogan, Harth, A., Umbrich, J., Kinsella, S., Polleres, A., & Decker, S., "Searching and browsing Linked Data with SWSE: The semantic web search engine," *Web Semantics: Science, Services and Agents on the World Wide Web,* pp. 365-401, 2011.

[27] R. Delbru, Stephane Campinas, and Giovanni Tummarello, "Searching web data: An entity retrieval and high-performance indexing model," *Web Semantics: Science, Services and Agents on the World Wide Web,* pp. 33-58, 2012.

[28] H. Wang, Qiaoling Liu, Thomas Penin, Linyun Fu, Lei Zhang, Thanh Tran, Yong Yu, and Yue Pan, "Semplore: A scalable IR approach to search the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web,* pp. 177-188, 2009.

[29] X. Dong, Alon Halevy, "Indexing dataspaces," presented at the international conference on Management of data (ACM SIGMOD), 2007.

[30] V. Lopez, Enrico Motta, Victoria Uren, "PowerAqua: Fishing the Semantic Web," presented at the The Semantic Web: Research and Application, 2006.

[31] D. Damljanovic, Milan Agatonovic, and Hamish Cunningham, "FREyA: An interactive way of querying Linked Data using natural language," presented at the In The Semantic Web: ESWC 2011 Workshops, 2012.

[32] A. Freitas, João Oliveira, Seán O'Riain, Edward Curry, João Pereira da Silva, "Querying Linked Data using semantic relatedness: A vocabulary independent approach," *Natural Language Processing and Information Systems,* pp. 40-51, 2011.

[33] A. Freitas, Joao Gabriel Oliveira, Edward Curry, Seán O'Riain, "A Multidimensional Semantic Space for Data Model Independent Queries over RDF Data," presented at the Semantic Computing (ICSC), Fifth IEEE International Conference on, 2011.

[34] O. Hartig, Juan Sequeda, Jamie Taylor, Patrick Sinclair, "How to consume Linked Data on the web: tutorial description," presented at the In Proceedings of the 19th international conference on World wide web, 2010.

[35] S. Bechhofer, Frank Van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, Lynn Andrea Stein, "OWL web ontology language reference," W3C recommendation 102004.

[36] D. Community. (2012/11/08). *http://dbpedia.org*.