British Journal of Mathematics & Computer Science
13(1): 1-14, 2016, Article no.BJMCS.21706
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Mining Semantic Web Data Using K-means Clustering Algorithm

Wria Mohammed Salih Mohammed[1] and Mohamad Mehdi Saraee[2*]

[1]School of Basic Education, Computer Science, University of Sulaimani, Kurdistan, Iraq.
[2]Islamic Azad University, Shahreza Branch, Shahreza, Islamic Republic of Iran.

Article Information

DOI: 10.9734/BJMCS/2016/21706
Editor(s):
(1) Doina Bein, Applied Research Laboratory, The Pennsylvania State University, USA.
Reviewers:
(1) G. Y. Sheu, Chang-Jung Christian University, Taiwan.
(2) M. Bhanu Sridhar, GVP College of Engineering for Women, Visakhapatnam, India.
Complete Peer review History: http://sciencedomain.org/review-history/12392

Original Research Article

Received: 30 August 2015
Accepted: 13 October 2015
Published: 21 November 2015

Abstract

The combination between semantic web and web mining is known as semantic web mining. Semantic web can improve the effectiveness of web mining. The knowledge of semantic web data can be mined using web mining techniques, as semantic web data are rich sources of knowledge to feed data mining techniques. This paper concentrated on how to combine two emergency research areas, namely semantic web and web mining. Firstly, we extract data from RDF file using SPARQL as query language. After that, we are going to cluster the entities of semantic web. One of the techniques is k-means clustering algorithm. Sematic web is about the meaning of the web data and to make machine understandable about it. Moreover, web mining is to extract and discover useful and previously unknown information from web data. This research gives an overview of where semantic web and web mining areas meet today, and how it is useful to combine these two well-known areas to obtain better and more accurate results.

Keywords: Data mining; semantic web; cluster; k-means.

1 Introduction

Over the past decade, semantic web and web mining have developed astonishingly. Semantic web has also received much attention from web community, and huge amount of work aimed to do on semantic web. Most of the web data are unstructured and humans cannot understand them. However, it can be seen that data can be processed efficiently by machines. Furthermore, in a simple way, semantic web is a web of data described and established context by linking that observe to explain grammar and language concepts [1]. In other words, as indicated in [2] "The Semantic Web is a mesh of information linked up in such a way as to

**Corresponding author: E-mail: saraee@iaush.ac.ir;*

be easily processable by machines". Web mining can extract and find useful and previously unknown information from web data [3]. From web, semantic web can make data machine-understandable.

The purpose of this paper is to mine semantic web by firstly, using RDF file document; secondly, using SPARQL as a query language to extract information from the RDF, and finally, mining the RDF file by using k-means cluster algorithms. The mixture between sematic web and web mining will greatly raise the understandability of the web for machines. As a result, we will see the benefit to use both well-known areas of sematic web and web mining to have better and more accurate results.

2 Related Work

Since over the past decades many approaches have been proposed for web mining and semantic web, it can be seen that there are many researches about semantic web and web mining and some other researches about mining semantic web used association and classification algorithms [4-7]. In [8] the authors, identified cluster semantic web into groups according to similarity and classified data-set into subsets or clusters. Furthermore, they [8] proposed two approaches to cluster semantic web resources; firstly, clustering an RDF based on instances, secondly, determining the distance between two instances. Likewise, in [9] they illustrated semantic web mining in general and focused on two main well-known areas; semantic web and web mining. However, in this paper they did not use a specific method. Berlanga and Nebot [10] presented works on semantic web to find association rules on it, using Novel method. In [7] they classified semantic web challenges, and then they used data mining technique on semantic web data, using ontology engineering.

Another research that published was about combining both semantic web and data mining, the main idea of this research is converting unstructured data into machine understandable data utilizing semantic web techniques, as a result, machine can answer to human questions in shorter time and avoid dull work, also, this paper particularly focus on semantic web techniques [11]. Furthermore, they [11] mentioned three different types of web mining areas, web content mining, web structure mining and web usage mining, finally, semantic based on web mining is applied in this research as well.

In [12] researchers found the benefits and drawbacks of K-means algorithm and then they combined K-means algorithm with semantic web ontology to create a semantic web model, however, they attempt to expand the existing K-means algorithm to make it more efficiency and precision in semantic web. In this paper, we are going to mine semantic web data using k-means clustering algorithms after applying SPARQL on RDF file.

3 Preliminaries

This section briefly describes semantic web and k-means clustering techniques which are related to our research area.

3.1 Semantic Web

Nowadays, there are many researches about semantic web. For example, Berners-Lee in [13] clarified semantic web as he said "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". It can be seen that semantic web is a web of data [14,15,13]. Furthermore, Berners-Lee suggested that semantic web is to make the web machine-processable [9]. The major building blocks of the Semantic Web are:

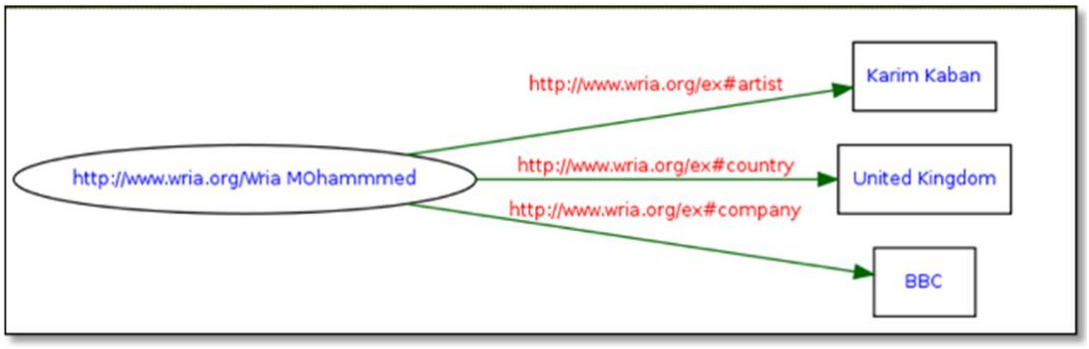
3.1.1 XML

XML stands for Extensible Markup Language. The main purpose of XML is to add markup to data. Furthermore, XML data can be transferred among users or applications [16].

3.1.2 RDF

RDF stands for Resource Description Framework, and it was created by W3C in 1999 [17]. RDF is a language for representing and describing information about resources on the web. It is specially used for representing metadata about resources on the web [18]. Moreover, it is used for exchanging machine-understandable information on the web [17]. Moreover, we have RDF/XML data such as:

```
<?xml version="1.0"?>
<RDF:RDF  xmlns:RDF="http://www.w3.org/1999/02/22-RDF-syntax-ns#"
          xmlns:ex="http://www.wria.org/ex#">
<RDF:Description
RDF:about="http://www.wria.org/Wria Mohammmed">
  <ex:artist>Karim Kaban</ex:artist>
  <ex:country>United Kingdom</ex:country>
  <ex:company>BBC</ex:company>
</RDF:Description>
</RDF:RDF>
```



Graph 1. RDF/XML example

3.1.3 Web ontology language

RDF-schema describes the resources using classes, and class properties. It means RDF-S allows developers to define the resources vocabulary. It can be seen that OWL is the extension of RDF-S [19]. In [20] they defined OWL as follows: “An OWL-encoded web-distributed vocabulary of declarative formalism describing a model of a domain”. The goal of OWL is to define ontologies including properties, sub-properties, classes and subclasses [19].

3.1.4 SPARQL

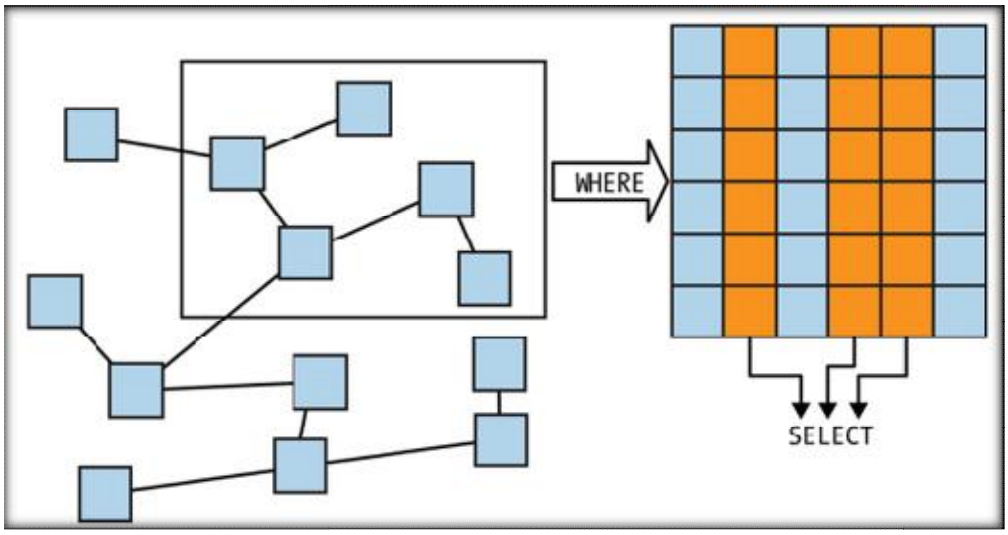
SQL is the most well-known query language for relational databases. However, the most important query language for semantic web is SPARQLS, which is a recommended query language for RDF by W3C [21]. The extension of SPARQL is (.rq) in lowercase. An example of SPARQL would be:

```
#FileName : book.rq
PREFIX ex: <http://www.wria.org/ex#>
SELECT ?primaryemail
WHERE
{ex:name    ex:email    ?primaryemail . }
```

This triple ended with period like Turtle. Moreover, it has subject ex:name, predicate ex:email and a variable is used instead of the object [22]. The analysis of the above example would be:

SELECT: It identified the appeared variable when the query was run (i.e. ?primaryemail).

WHERE clause includes one pattern (ex:name ex:email ?primaryemail). And it matched with the triple inside RDF file which this SPARQL was applied on [21].



Graph 2. WHERE specifies to extract data, and SELECT chooses data to display [22]

3.2 Data Mining Clustering Approach

Clustering is the classification of a data objects into subsets. Each subset is called a cluster. All elements in the same cluster are similar to each other and different from other clusters. For this technique, clustering algorithms are used and humans do not perform it. Clustering is unsupervised learning; it means clustering is useful to find the previous unknown clusters in the data [23].

Clustering approach has four main methods to cluster data:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods

Partitioning methods have some algorithms to cluster dataset; one of them is k-means algorithm.

In this paper we are going to focus k-means algorithms, which is one of the most well-known and oldest algorithms [24].

K-means is one of the point-assignment algorithms to cluster a data set to subsets. Firstly, to apply k-means algorithms, it needs the number of clusters (K); secondly, select the points (centroids) far from each other. To make cluster, the closest points from a centroid will be in the same cluster. Euclidean space or Manhattan distance is used to measure the distance between centroids and all points [25]. There are some brief steps to apply K-means algorithms:

1- Choosing the number of clusters (k).
2- Picking the centroids. The number of centroids is equal to the number of clusters. The centroids will be chosen randomly.
3- Measuring the distance between centroids with each object.
4- Making each object an element of a group according to the distance between centroids and the element.
5- Checking whether the process has finished or restarted from step 3 [26].

When we apply k-means, we will need:

- Dataset containing objects (D)
- Selecting the number of clusters (k)

The result will be a set of k clusters [23].

4 Problem Statement

The problem is to implement data mining technique on semantic web using SPARQL to retrieve data from RDF file, and to apply K-means algorithms to mine the datasets. This paper is concerned with k-means clustering algorithm and query language (SPARQL) on RDF file. A K-means algorithm is a technique to cluster objects into groups. Furthermore, k-means clustering approach is unsupervised learning. There are some clusters examples:

- Clustering web usage pattern according to similarity.
- Grouping cars into categories according to similar size [27].

SPARQL querying language has been developed amazingly in parallel with RDF and ontology and there are some researches about it. In this paper we are using SPARQL to retrieve data from RDF file.

All data items in RDF file are shown in the format of triples (subject, predicate and object); for example, information about (book) could include triples [28]:

Subject	Predicate	Object
http://www.wria.org/	http://wria.info/book1.1title	"clustering semantic web mining"
http://www.wria.org/	http://wria.info/book1.1author	"Professcr.Saraee"
http://www.wria.org/	http://wria.info/book1.1edition	"1"

It can be seen that RDF file can be represented and created in several ways. This increases testing and making different query language on it. However, nowadays different query mechanisms can be implemented on RDF file [29].

5 Methodology

Mining semantic web using SPARQL and K-means algorithm is implemented using some different applications. This approach uses SPARQL to obtain dataset from RDF file and after that applies K-means algorithm to mine it. How these two methods work together can be summarized as follows:

> There are many RDF files or OWL files that we can implement on it. In this research an RDF file is used (book.owl) which was created for this paper. This example can be created in any simple text editor such as notepad. RDF has triples that consist of three main parts: Subject, predicate and object.

Writing SPARQL as a query language for RDF file, the purpose of using SPARQL is to extract information from RDF file. As a result, we will have a table to make it easy to mine. It is clear that to use SPARQL, it needs a server (such as JENA apache). In this paper, we use online tool to write SPARQL (http://demo.openlinksw.com/sparql/) and then we get dataset. After that, we can save the result as a CSV file.

As a result, we will have the table of data including entities such as URI or literal. Finally, we can mine it using k-means technique algorithm.

5.1 Experimental Step

This approach was developed and tested using a dataset created for this purpose. Here is a sample of the XML/RDF file:

```
<?xml version="1.0" encoding="UTF-8"?>
<RDF:RDF xmlns:RDF="http://www.w3.org/1999/02/22-RDF-syntax-ns#"
 xmlns:ex="http://example.org/">
          <RDF:Description RDF:about="http://example.org/book">
            <ex:title>
              <RDF:Description RDF:about="http://example.org/Iris Disease Classifying ">
                    <ex:author RDF:resource="http://example.org/Mohammad Mehdi Saraei" />
                    <ex:year>2009</ex:year>
                    <ex:publisher RDF:resource="https://www.springer.com/Springer" />
                    <ex:ISBN>978-3-642-01215-0</ex:ISBN></RDF:Description>
      </ex:title>
    </RDF:Description>
         <RDF:Description RDF:about="http://example.org/book">
           <ex:title>
      <RDF:Description RDF:about="http://example.org/A New Linear Appearance-based
      Method in Face Recognition ">
             <ex:author RDF:resource="http://example.org/Mohammad Mehdi Saraei" />
             <ex:year>2008</ex:year>
             <ex:publisher RDF:resource="https://www.springer.com/Springer" />
             <ex:ISBN>978-0-387-74937-2</ex:ISBN></RDF:Description>
      </ex:title>
    </RDF:Description>
                  .
          </RDF:RDF>
```

As it can be seen, the element values consist of linked data and nominal literals. i.e. *http://example.org/Mohammad Mehdi Saraei* is an URI record.
And <ex:year>2009</ex:year> it is a literal data.

5.2 Experimental Results

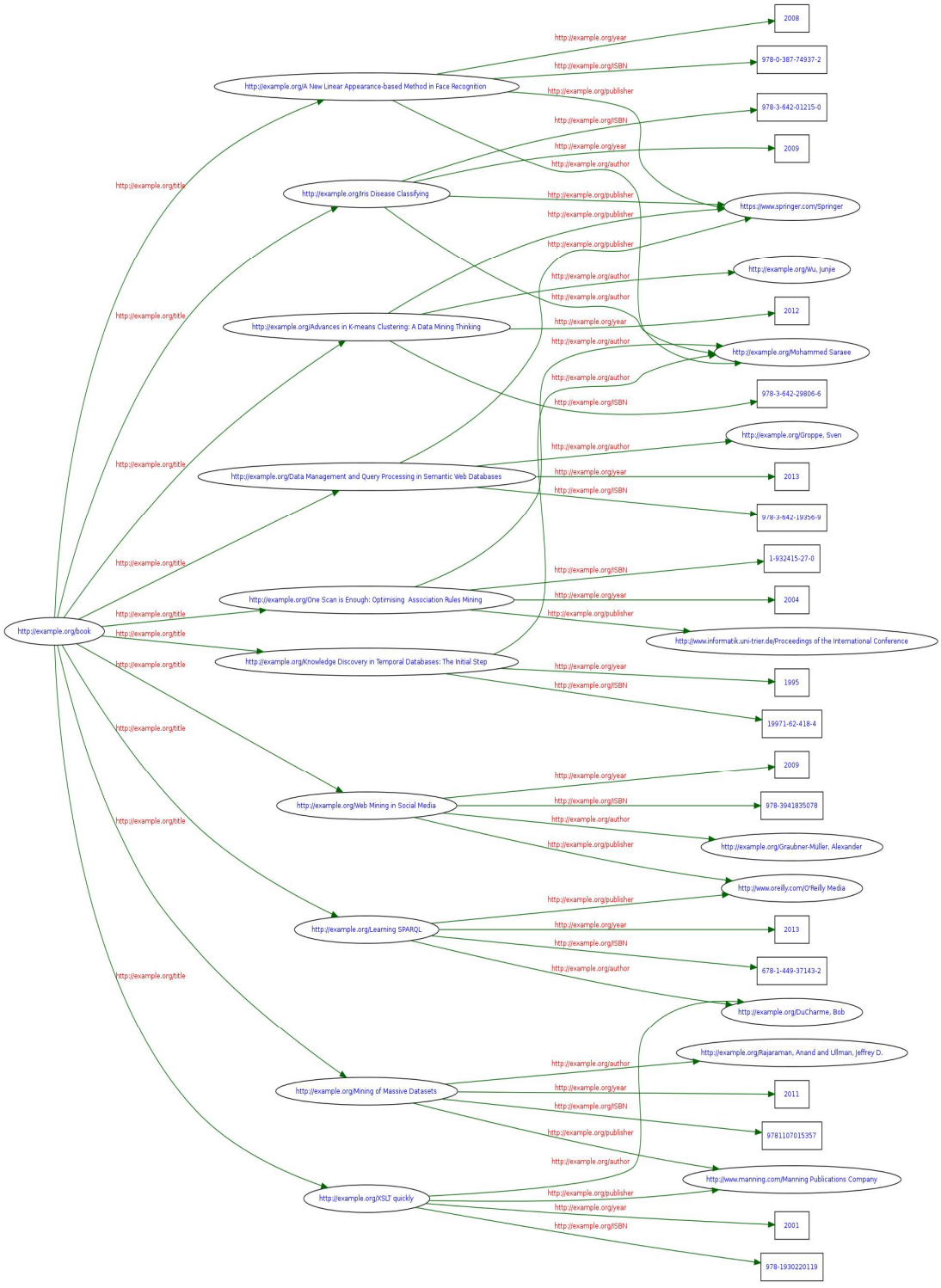
Firstly, to create an example of RDF file and then validate it using http://www.w3.org/RDF/Validator/), the following results is presented:

- Triple of the data model:

Table 1. Triples (subject predicate and object)

Subject	Predicate	Object
http://example.org/book	http://example.org/title	http://example.org/Iris Disease Classifying
http://example.org/Iris Disease Classifying	http://example.org/author	http://example.org/Mohammed Saraee
http://example.org/Iris Disease Classifying	http://example.org/year	"2009"
http://example.org/Iris Disease Classifying	http://example.org/publisher	https://www.springer.com/Springer
http://example.org/Iris Disease Classifying	http://example.org/ISBN	"978-3-642-01215-0"
http://example.org/book	http://example.org/title	http://example.org/A New Linear Appearance-based Method in Face Recognition
http://example.org/A New Linear Appearance-based Method in Face Recognition	http://example.org/author	http://example.org/Mohammed Saraee
http://example.org/A New Linear Appearance-based Method in Face Recognition	http://example.org/year	"2008"
http://example.org/A New Linear Appearance-based Method in Face Recognition	http://example.org/publisher	https://www.springer.com/Springer
http://example.org/A New Linear Appearance-based Method in Face Recognition	http://example.org/ISBN	"978-0-387-74937-2"
http://example.org/book	http://example.org/title	http://example.org/Web Mining in Social Media
http://example.org/Web Mining in Social Media	http://example.org/author	http://example.org/Graubner-Müller, Alexander
http://example.org/Web Mining in Social Media	http://example.org/year	"2009"
http://example.org/Web Mining in Social Media	http://example.org/publisher	http://www.oreilly.com/O'Reilly Media
http://example.org/Web Mining in Social Media	http://example.org/ISBN	"978-3941835078"
http://example.org/book	http://example.org/title	http://example.org/Mining of Massive Datasets

- The graph for this example will be the following:



Graph 3. The data model

Secondly, after using SPARQL, the dataset from RDF can be obtained, and then save it as CSV file. As a result, we can use CSV file in WEKA application or SAS Enterprise Miner.

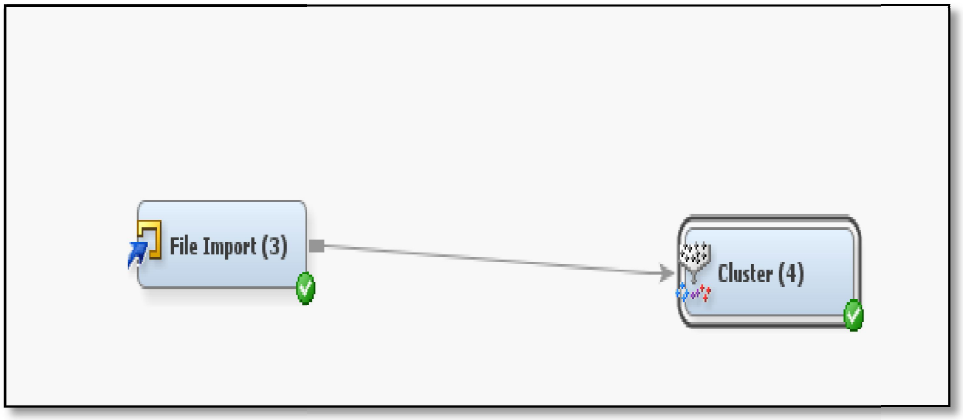
name	author	year	publisher	ISBN
http://example.org/A New Linear Appearance-based Method in Face Recognition	http://example.org/Mohammed Saraee	2008	https://www.springer.com/Springer	978-0-387-74937-2
http://example.org/Advances in K-means Clustering: A Data Mining Thinking	http://example.org/Wu, Junjie	2012	https://www.springer.com/Springer	978-3-642-29806-6
http://example.org/Data Management and Query Processing in Semantic Web Databases	http://example.org/Groppe, Sven	2013	https://www.springer.com/Springer	978-3-642-19556-9
http://example.org/Iris Disease Classifying	http://example.org/Mohammed Saraee	2009	https://www.springer.com/Springer	978-3-642-01215-0
http://example.org/Knowledge Discovery in Temporal Databases: The Initial Step	http://example.org/Mohammed Saraee	1995	http://www.ibo.org/International Conference and Workshops	19971-62-418-4
http://example.org/Learning SPARQL	http://example.org/DuCharme, Bob	2013	http://www.oreilly.com/O'Reilly Media	978-1-449-37143-2
http://example.org/Mining of Massive Datasets	http://example.org/Rajaraman, Anand and Ullman, Jeffrey D.	2011	http://www.manning.com/Manning Publications Company	9.78111E+12
http://example.org/One Scan Is Enough: Optimising Association Rules Mining	http://example.org/Mohammed Saraee	2004	http://www.informatik.uni-trier.de/Proceedings of the International Conference	1-932415-27-0
http://example.org/Web Mining in Social Media	http://example.org/Graubner-MÃ¼ller, Alexander	2009	http://www.oreilly.com/O'Reilly Media	978-3941835078
http://example.org/XSLT quickly	http://example.org/DuCharme, Bob	2001	http://www.manning.com/Manning Publications Company	978-1930220119

Graph 4. Extraction entities from RDF file

5.3 K-means Clustered Results

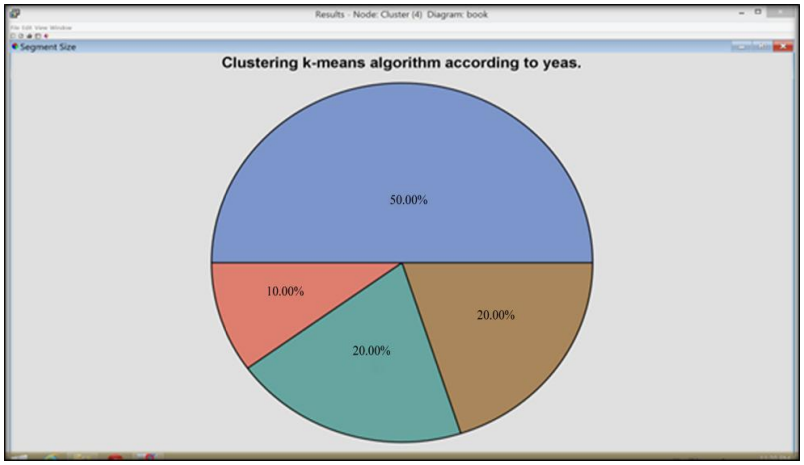
To mine the dataset we have, we used three different mining tools to cluster the dataset using k-means algorithms.

1- SAS enterprise miner, by adding cluster nodes on the dataset.



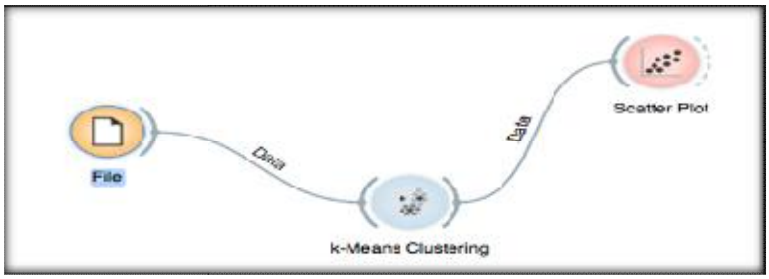
Graph 5. Clustering in SAS enterprise miner

And we will have the following clustering according to years, we clustered according to the year of the publication and we will have the following result:



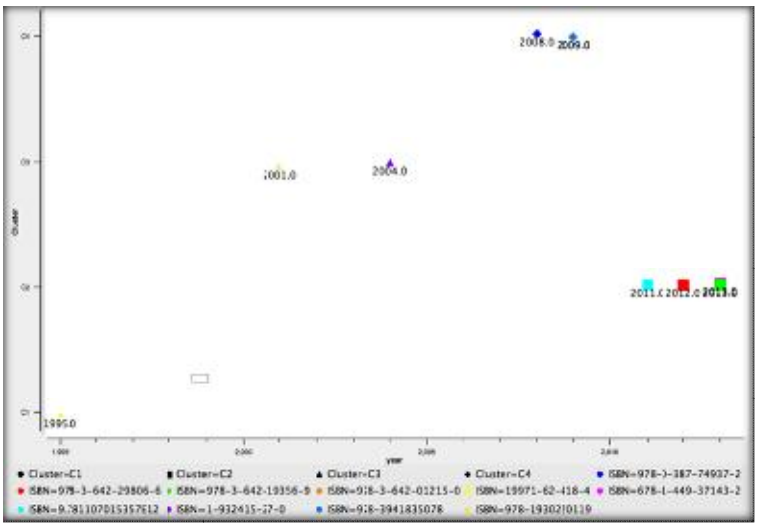
Graph 6. Pie chart of 4 clusters

2- Orange data mining: we applied k-means algorithm using Orange data mining application



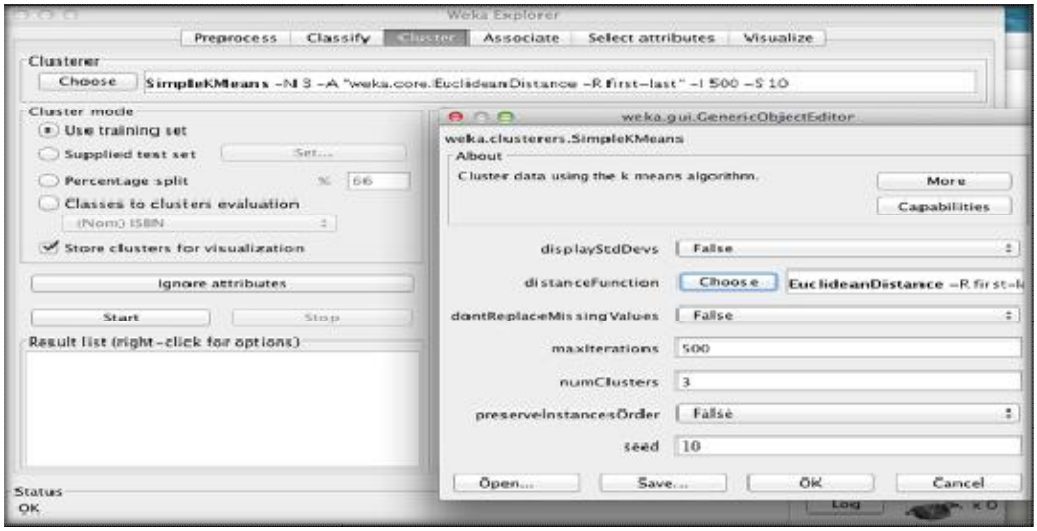
Graph 7. K-Means using orange application

The following result will be presented:



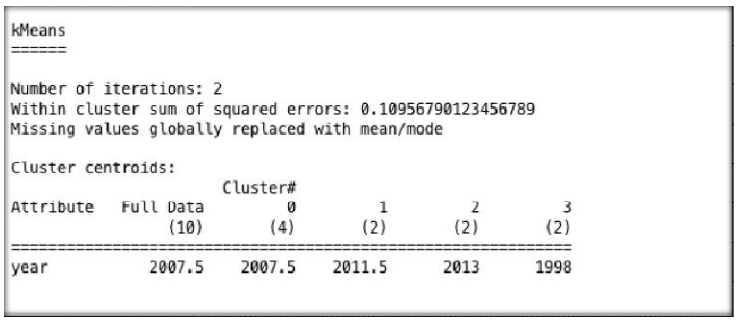
Graph 8. The clusters in Orange application

3- Then we used Weka to mine the data:



Graph 9. K-Means using Weka

The following result will be showed:



```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 0.10956790123456789
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute   Full Data         0         1         2         3
                 (10)       (4)       (2)       (2)       (2)
=============================================================
year          2007.5    2007.5    2011.5      2013      1998
```

Graph 10. Weka application clustering techniques

6 Discussion

This research is finding the influences of the combination between semantic web and data mining. First of all, dataset is extracted from semantic web data using SPARQL, as a result cleaned dataset can be obtained from semantic web data, and it is ready to be mined. Second of all, K-means clustering algorithm is applied on cleaned dataset, which is extracted from semantic web data, finally, there is several clusters. The results in Table 1 shows that there are triples (subject, predicate and object) of the RDF file which contain URI or literals. Furthermore, literals are numbers or string. Moreover, graph 1 is a graph of the RDF/XML file that illustrates the relationships among entities. After that, we extracted data from RDF file using SPARQL as a query language. We can see the result of database table in graph 4 that includes all data from RDF file. From graph 4, it can be noticed that all necessary dataset can be extracted from semantic web data. Finally, the extracted dataset from semantic web data can be mined using k-means clustering algorithm to 4 groups which are appeared in graph 6 as a pie chart, graph 8 shows four different groups and graph 10 shows 4 different centroids. Furthermore, in this research different tool are used such as Weka, orange and SAS enterprise miner to obtain efficient and accurate results.

7 Conclusion

In this paper the application of data mining techniques to mine interesting patterns from semantic web data has been presented. Our proposed data mining technique is clustering and in particular K-means algorithm is used to cluster semantic web data. Clustering techniques are mostly unsupervised methods that can be used to organize data into groups based on similarities among the individual data items. RDF file that includes (subject, predicate and object), have been used to evaluate and SPARQL employed to query RDF data. K-means has generated a number of good clusters which clearly shows the suitability of the technique for Semantic Web data. In conclusion, it can be clearly seen that the combination between semantic web and data mining can obtain a efficiency results, particularly, k-means clustering algorithm is one of the most well-known technique that can divide the semantic web data into several groups, as a result, each group can be understandable easily. As a future work, we intend to apply fuzzy clustering algorithm on semantic web data since this method allows objects to belong to several clusters simultaneously with different degrees of membership.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez. Semantic web programming. s.l.: John Wiley & Sons; 2011.

[2] Palmer Sean B. Sean B. Palmer. [Online] September 2001. [Cited: April 06, 2014]. Available: www.infomesh.net

[3] Graubner-Müller, Alexander. Web mining in social media: Use cases, business value, and algorithmic approaches for corporate intelligence. s.l.: Social Media Verlag. 2011;978-3941835078.

[4] Anyanwu Kemafor, Sheth Amit. *P*-Queries: enabling querying for semantic associations on the semantic web. New York: WWW '03 Proceedings of the 12th International Conference on World Wide Web. 2003;690-699:1-58113-680-3 .

[5] Dave Kushal, Lawrence Steve. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. New York: WWW '03 Proceedings of the 12th International Conference on World Wide Web. 2003;1-58113-680-3.

[6] Patel Chintan, et al. OntoKhoj: a semantic web portal for ontology searching, ranking and classification. New York, USA: WIDM '03 Proceedings of the 5th ACM International Workshop on Web Information and Data Management. 2003;1-58113-725-7.

[7] Keyvanpour MohammadReza, Hassanzadeh Hamed, Khoshroo Babak Mohammadizadeh. Comparative classification of semantic web challenges and data mining techniques. s.l.: International Conference on Web Information Systems and Mining; 2009.

[8] Grimnes Gunnar Aastrand, Peter Edwards, Alun Preece. Instance based clustering of semantic web resources. Springer Berlin Heidelberg, The Semantic Web: Research and Applications. 2008;303-317.

[9] Shukla Arti, Priyanka Yadav. Semantic web mining: Review. 12, Chicago: International Journal on Recent and Innovation Trends in computing and communication. 2013;1:232-1869-919-922.

[10] Nebot Victoria, Rafael Berlanga. Finding association rules in semantic web data. Castellón, Spain: Knowledge-Based Systems. 2012;51–62.

[11] Semantic web mining - A review. 12, s.l.: International Journal of Computer Applications (0975 – 8887). 2015;117 :12.

[12] Wentian Guo Qingju Ji, Sheng Zhong, En Zhou. The analysis of the ontology-based K-means clustering algorithm. Paris, France: Atlantis Press. International Conference on Computer Science and Electronics Engineering. 2013;734-737.

[13] www.w3.org. [Online] W3C, May 2001. [Cited: April 03, 2014]
Available: http://www.w3.org/RDF/Metalog/docs/sw-easy

[14] Benjamins V. Richard, et al. Law and the Semantic web. New York: Springer Berlin Heidelberg; 2004;3-540-25063.

[15] Kashyap Vipul, Bussler Christoph, Moran Mathew. The semantic web. New York: Springer-verlag Berlin Heidelberg. 2008;978-3-540-76451-9.

[16] Jim Melton, Stephen Buxton. Querying XML: XQuery, XPath, and SQL/XML in context. United State of America: Morgan Kaufmann. 2011;978-1-55860-711-8.

[17] Yu Liyang. A Developer's guide to the semantic web. New York: Springer Heidelberg Dordrecht London. 2011;978-3-642-15969-5.

[18] Straccia Umberto. Foundations of fuzzy logic and sematic web languages. Pisa, Italy: CRC Press. 2014; 978-1-4398-5347.

[19] Sajja Priti Srinivas, Akerkar Rajendra. Intelligent technologies for web applications (Chapman & Hall/CRC data mining and knowledge discovery series). Minnesota USA: Chapman and Hall/CRC. 2012;978-1439871621.

[20] Lacy Lee W. OWL: Representing information using the web ontology language. s.l.: Trafford Publishing. 2006;978-1412034487.

[21] Groppe Sven. Data management and query processing in semantic web databases. New York: Springer. 2011;978-3-642-19356-9.

[22] DuCharme Bob. Learning SPARQL. s.l.: O'Reilly Media. 2013;678-1-449-37143-2.

[23] Han Jiawei, Kamber Micheline, Pei Jian. Data mining: Concepts and techniques. Waltham USA: Morgan Kaufmann. 2012;978-0123814791.

[24] Wu Junjie. Advances in K-means clustering: A data mining thinking. New York: Springer. 2012; 978-3-642-29806-6.

[25] Rajaraman Anand, Ullman Jeffrey D. Mining of massive datasets. s.l.: Cambridge University Press. 2011;9781107015357.

[26] Srija Unnikrishnan, Sunil Surve, Deepak Bhoir. Advances in computing, communication and control. Mumbia, India: Springer. 2011;978-3-642-18440-6.

[27] Saraee Mohammed, Aljibouri Joanna Moaiad. Mining XML data: A clustering approach. Las Vegas, USA: CSREA Press; 2005.

[28] Neumann Thomas, Weikum Gerhard. The RDF-3X engine for scalable management of RDF data. Saarbrucken, Germany: s.n.; 2009.

[29] Gutierrez Claudio, Hurtado Carlos, Mendelzon Alberto O. Foundations of semantic web databases. s.l.: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. ACM; 2004.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/12392