Mining Semantic Web Data using K-Means

Clustering Algorithm

Paper By: Wria Mohammed Salih Mohammed

And

Mohamad Mehdi Saraee

Digital Assignment

By:

Shashanka Shekhar Das

17BCE1325

**Abstract:**

The underlying idea behind this paper is to combine semantic web and web mining to produce a hybrid called semantic web mining. The paper claims that semantic web improves the efficacy of web mining. The paper uses SPARQL as the query language to extract data from RDF files and uses clustering algorithms like k-means clustering to mine the data from the RDF file.

**Semantic Web:**

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

**RDF:**

RDF stands for Resource Description Framework and created by W3C in 1999. RDF is a language for representing and describing information about resources on the web, especially used for representing metadata about resources on the web.

**SPARQL:**

SPARQL is a query language and protocol designed for RDF by W3C RDF Data Access Working group. It is a lot similar to SQL but used only in RDF.

**Problem Statement:**

The problem is to implement data mining technique on semantic web using SPARQL to retrieve data from RDF file and then to apply K-means algorithm to mine the dataset.

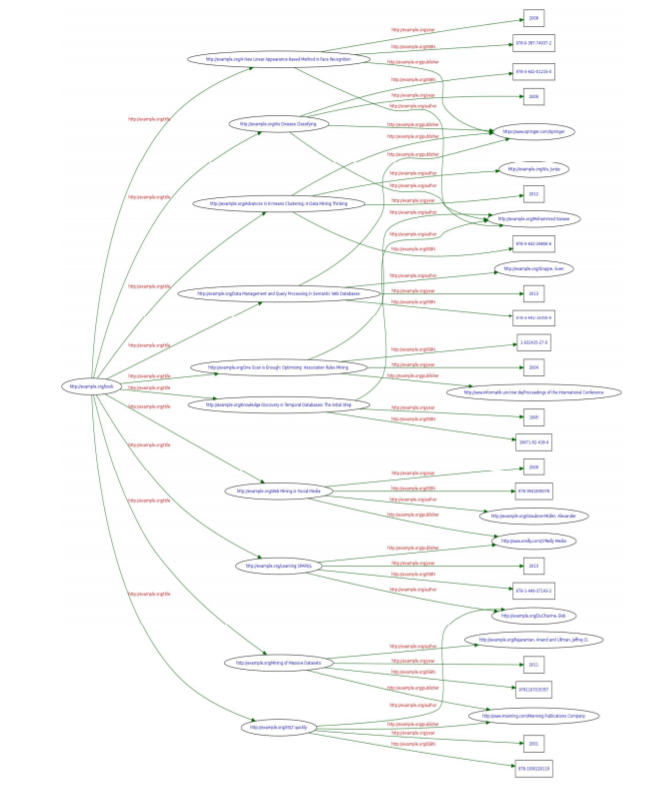
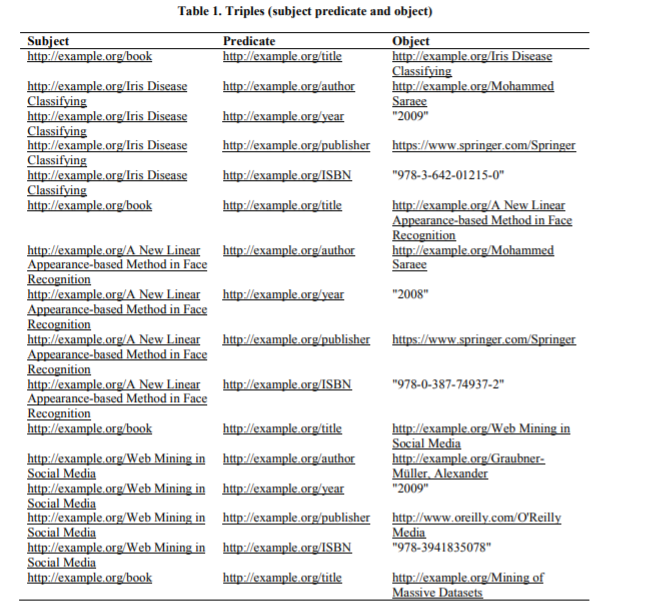
**Methodology:**

The paper uses an approach which uses SPARRQL to obtain dataset from the RDF file and then applies K-means algorithm to mine it.

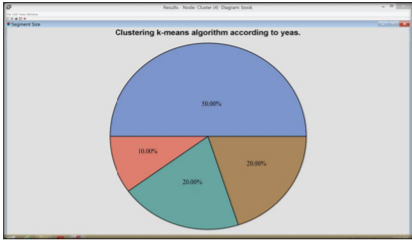
This paper used an RDF file (book.owl), they specifically created this file for this paper. RDF has triples that usually consists of three main parts: Subject, predicate and object.

To use SPARQL one needs a server, they used (http://demo.openlinksw.com/sparql/) and then after procuring the dataset and querying it, they save the result in the form of a CSV file. Then use K-mining to form previously unknown clusters.

**Used dataset:**



*Rdf file (book.owl) Graph for the corresponding dataset*



*After using SPARQL to extract data from RDF file K-means clustering on given dataset*