LiDDM: A Data Mining System for Linked Data

Venkata Narasimha Pavan Kappara
Indian Institute of Information Technology
Allahabad, India
kvnpavan@gmail.com

Ryutaro Ichise
National Institute of Informatics
Tokyo, Japan
ichise@nii.ac.jp

O.P. Vyas
Indian Institute of Information Technology
Allahabad, India
opvyas@iiita.ac.in

Contents

- Background
- LiDDM: A Model
- Implementation Work
- Case Study
- Discussions and Future Work
- Conclusions

Background

- The quantity of linked data is growing rapidly
- Linking Open Data(LOD) forms the foundation for linking the data available on the web in structured format
- The result related to user query for extracting a useful hidden pattern may not always be completely answered by using only one or many of the datasets in isolation

Background

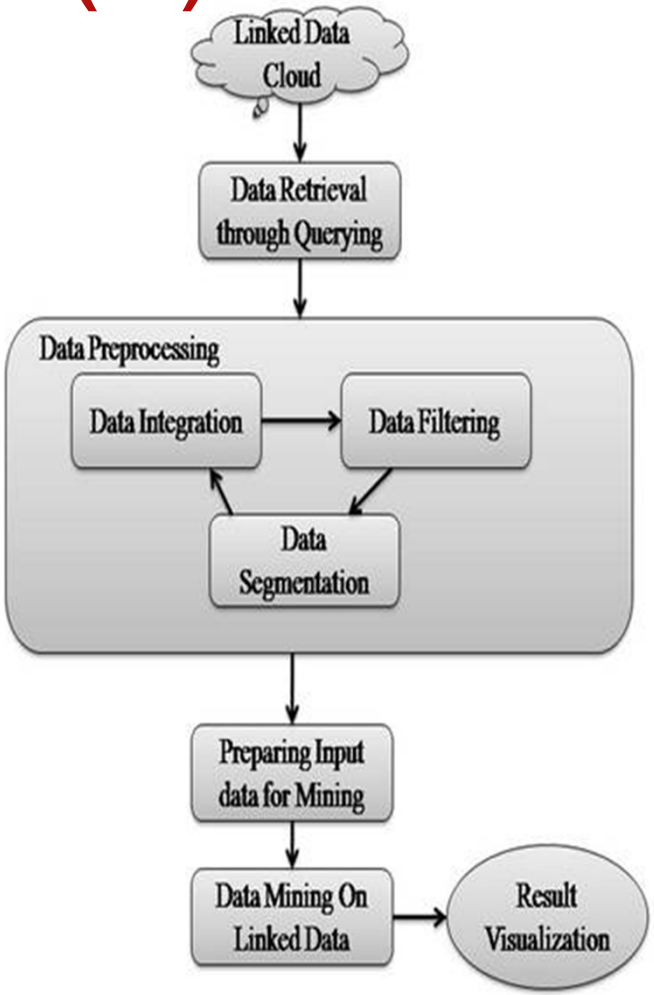
- Here linked data comes into picture as there is a need to integrate different data sources available in different structured formats to answer such type of complex queries
- Our model is targeted to deal with the complexities associated with mining the linked data efficiently

Approach

- Our hypothesis is implemented in form of a tool that
  - takes the data from linked data cloud,
  - performs various Knowledge Discovery in Databases (KDD) operations on linked data
  - applies data mining techniques such as association, clustering etc.
  - visualizes the result at the end.

LiDDM: A Model (1)

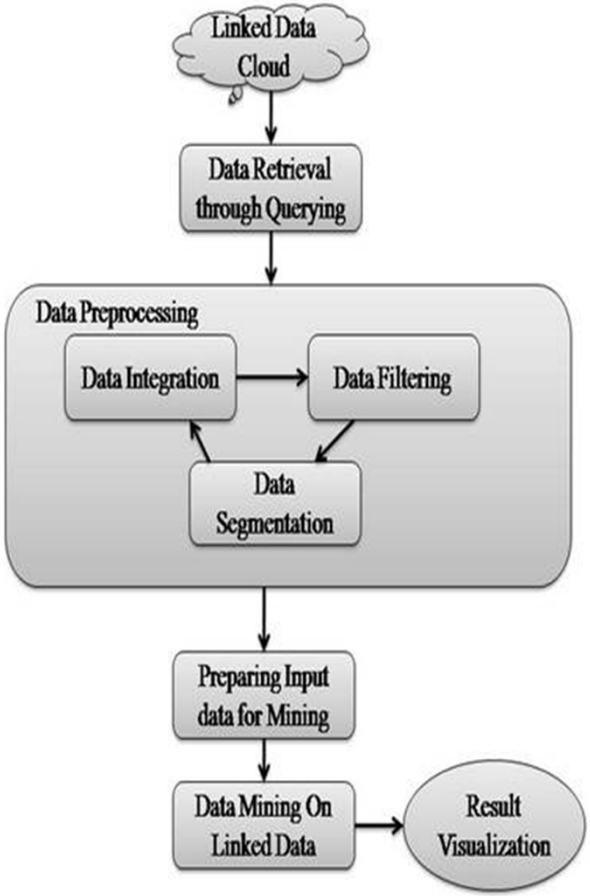
- Our model modified the process of KDD to conform to the needs of linked data and proceeded in a hierarchical manner.
- Here comes the different steps involved in it..



Architecture of LiDDM

LiDDM: A Model (2)

- Data Retrieval through Querying
  - This step can be compared to the data selection step in the KDD Process
- Data Pre-processing
  - This process has three sub steps. They are
    1. Data Integration
    2. Data Filtering
    3. Data Segmentation



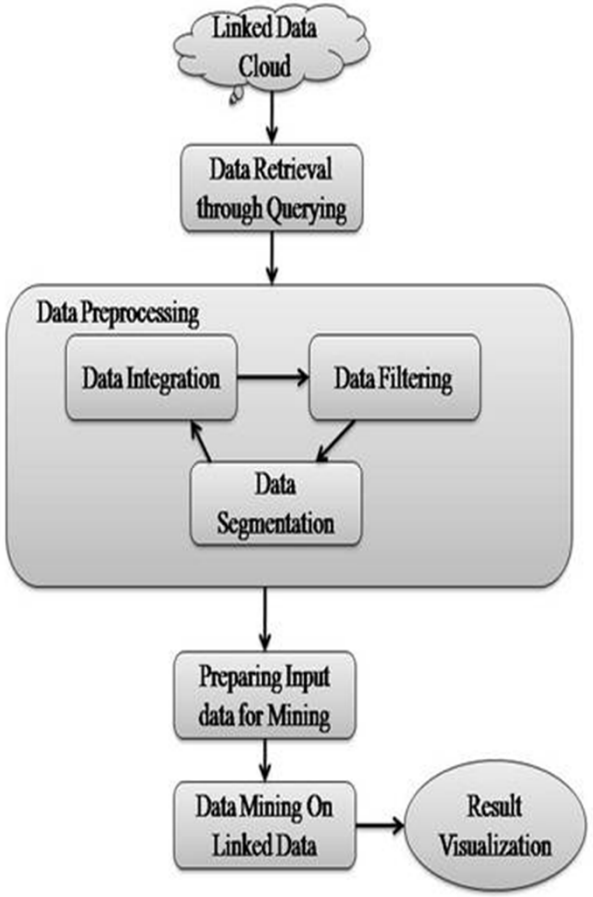
Architecture of LiDDM

LiDDM: A Model (3)

1. Data Integration:
   - Data is integrated based on some common relation presented in respected data sources.
   - Data sources are selected depending on different factors a user wants to study in different sources.
2. Data Filtering:
   - Data Filtering eliminates unwanted data and attributes from the integrated data and also constraints the data within some bounds.
3. Data Segmentation:
   - Data is classified into different classes and segments if needed.

LiDDM: A Model (4)

- Preparing Input Data for Mining
  - The format in which we retrieve the linked data has to be converted into a correct format that is required for feeding into the data mining system

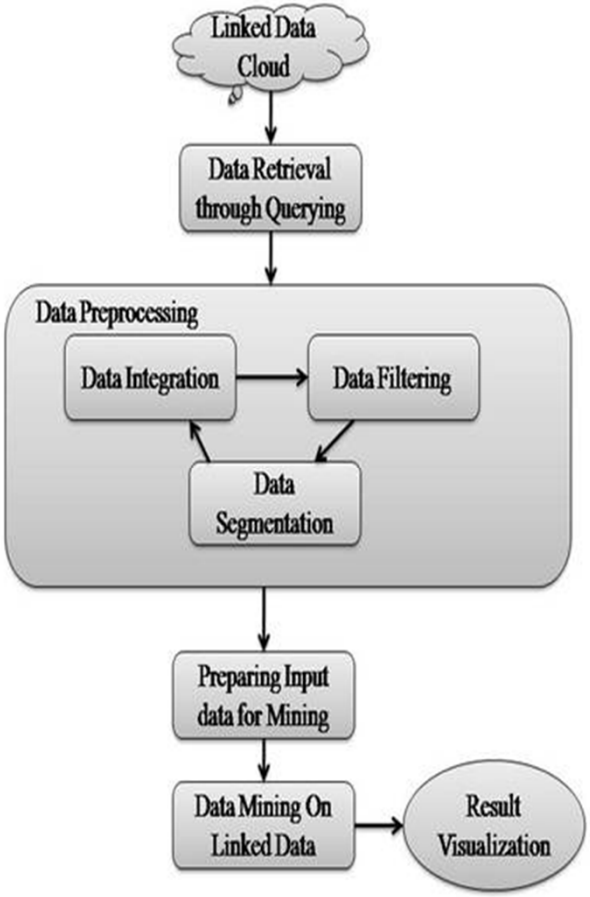


Architecture of LiDDM

LiDDM: A Model (5)

- Data Mining on Linked Data
  - Here the data may be classified or clustered or set for finding association rules.
  - The results are obtained and visualized for interpretation.

Thus LiDDM will ensure a very good and easy to use framework for interacting with Linked Data, reshaping and visualizing the results.



Architecture of LiDDM

IMPLEMENTATION WORK (1)

Step 1:

- Two types of querying are implemented.
  - One asks the user for a direct SPARQL Query and SPARQL end point.
  - The second one does an automatic query building and asks the user only for triples.

Step 2:

- Data integration can be done in two ways.
  - One way is performing a JOIN operation on the data retrieved.
  - The second way is to append both the results end to end if they have same data types.

IMPLEMENTATION WORK (2)

Step 3:

- UI has features of removing unwanted columns, deleting the rows that have values out of a certain range in a numerical column, deleting the rows that have certain strings in certain columns, etc.

Step 4:

- both numerical and string based segmentation is done.

Step 5:

- data is converted into ARFF(Attribute-Relation File Format) format for WEKA to work on it.

IMPLEMENTATION WORK (3)

Step 6:

- a separate UI for using original WEKA and a simplified UI are provided for quick mining.
- Simplified UI features J48 decision tree for classification , Apriori algorithm for association and EM(Estimation Maximization) for clustering.
  - The results from J48 decision tree are visualized in the form of a decision tree with precision, recall, F-Measure etc.
  - The results from EM clustering are visualized in the form of some clusters on the axes.

CASE STUDY (World FactBook)

- The first case study focuses on data from World FactBook.
  - World FactBook Database is queried for
    - GDP per capita
    - GDP composition by agriculture
    - GDP composition by industry
    - GDP composition by services of every country
  - Then Segmentation is done and the data is divided into different classes independently for each column.

CASE STUDY (World FactBook)

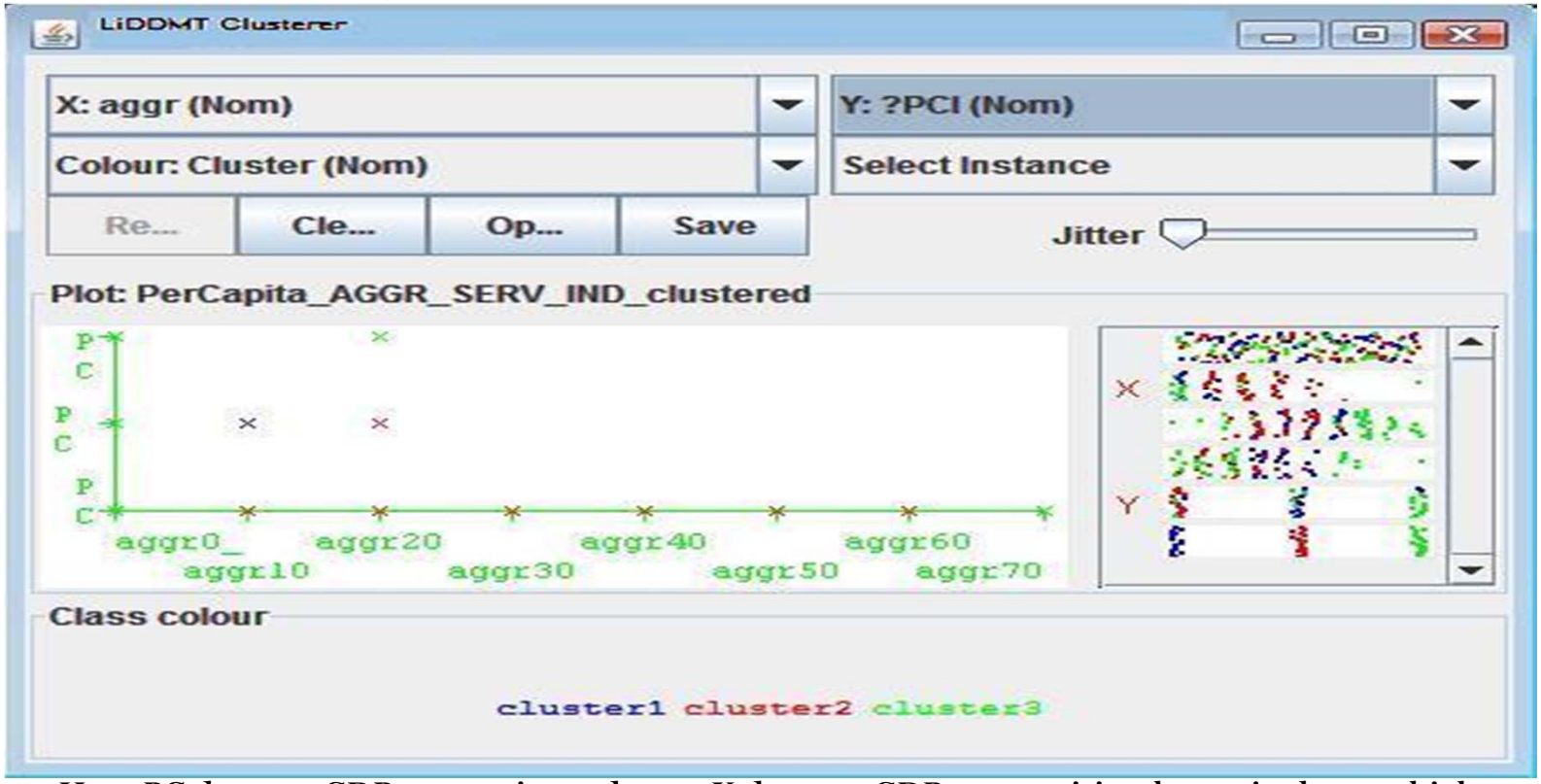
- Apriori Association gave the following output.

When the GDP per capita income is high (40 instances), the GDP composition by agriculture is between 0 to 10 percent (39 instances) with a confidence of 0.98.

When the GDP composition by services is between 70 to 80 percent (32 instances), the GDP composition by agriculture is between 0 to 10 percent (29 instances) with a confidence of 0.91

CASE STUDY (World FactBook)

- The same data is allowed to undergo EM clustering. The results also prove the same.



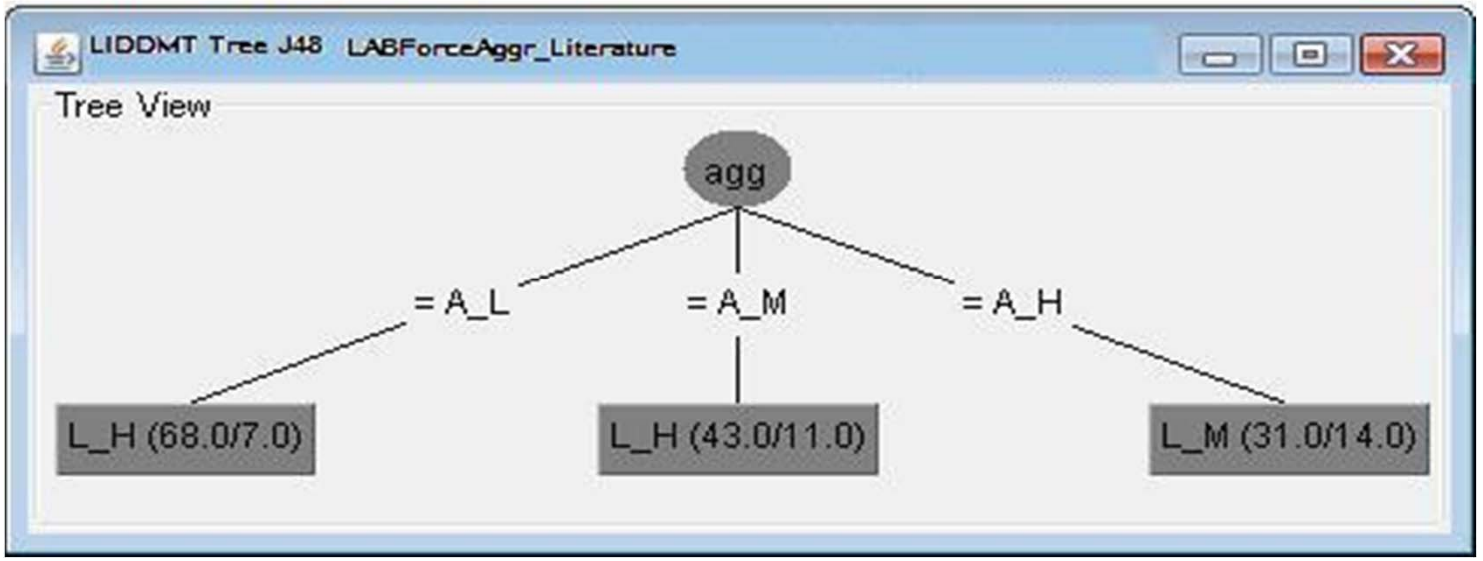
Here PC denotes GDP per capita and aggr-X denotes GDP composition by agriculture which is X percent

CASE STUDY (World FactBook)

- In order to analyze further, World FactBook is queried for
  - literacy rate,
  - labour force in agriculture
  - labour force in industry
  - labour force in services of every country.

CASE STUDY (World FactBook)

- Result of decision tree for predicting literacy rate



This figure shows that when labour force from agriculture is low (A L), then literacy rate is high (L H) with a 7 percent error rate out of 68 instances. Also when labour force from agriculture is medium (A M), then the literacy rate is high (L H) with 11 percent error rate out of 43 instances. Thus this can signify an inverse relationship between literacy rate and labour force in agriculture

CASE STUDY (Multiple Data)

- Information about movies from 1991 to 2001 by DBPedia and Linked Movie Data Base from various countries is retrieved
- The data is integrated with data retrieved from the World FactBook like
  - median age of the population
  - total population

CASE STUDY (Multiple Data)

- Our system found out the following patterns.

If the population is greater than 58,147,733 and median age is greater than 38, the movie production is high with a confidence of 1.

If population is between 58,147,733 and 190,010,647 and median age is less than 38, the movie production is low with a confidence of 1.

DISCUSSIONS

- Our model can treat data from various sources in the same way and also integrate them.
- Also our tool, LiDDMT, helps us to mine and visualize data from more than one SPARQL end point at the same time.

FUTURE WORK

- By introducing graph-based techniques, triples could be found out automatically in future.
  - More functionality can be imparted into the automatic query builder.
- Some artificial intelligence measures can be introduced into LiDDM for suggesting the best machine learning algorithms that can give the best possible results depending on the data obtained from the linked data cloud.

CONCLUSIONS

- There is a need to mine Linked Data to find different hidden patterns and also make it conceivable for people to find out what it has in store for us.
- Our model, LiDDM, successfully builds a data mining mechanism on top of linked data for effective understanding and analysis of linked data.
- The features in our model are built upon the classical KDD process and are modified to serve the needs of linked data.

CONCLUSIONS

- The step of getting the required data from the remote database itself makes our model dynamic.
- Using WEKA in our tool for the process of data mining makes it more efficient considering the vast popularity of WEKA.
- Also, having a chance to view more than one visualization at a time when implementing more than one data mining method makes our tool a very suitable one to compare data.