Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LINKED OPEN GOVERNMENT DATA AS BACKGROUND KNOWLEDGE IN PREDICTING FOREST FIRE

[1]GURUH FAJAR SHIDIK, [2]AHMAD ASHARI
[1]Universitas Dian Nuswantoro, Semarang, Indonesia
[2]Universitas Gadjah Mada, Yogyakarta, Indonesia
Email: [1]guruh.fajar@research.dinus.ac.id , [2]ashari@ugm.ac.id

ABSTRACT

Nowadays with linked open data, we can access numerous data over the world that more easily and semantically. This research focus on technique for accessing linked open government data LOGD from SPARQL Endpoint for resulting time series historical of Forest Fire data. Moreover, the data will automatically uses as background knowledge for predicting the *number of forest fire* and *size of burn area* with machine learning. By using this technique, LOGD could be used as an online background knowledge that provide time series data for predicting trend of fire disaster. In evaluation, mean square error MSE and root mean square error RMSE are used to evaluate the performance of prediction in this research. We also compare several algorithm such as Linear Regression, Neural Network and SVM in different window size.

Keywords: *Linked Open Government Data, Forest Fire Prediction, Time Series Data, Data Mining.*

1 INTRODUCTION

Forest fire is a common natural world phenomenon. Every year millions of hectares of forests in the worldwide are destroyed, in 1980 - 2007 at least 2.7 million hectares were burnt in Portugal [1]. This causes severe damages to the natural environment and results in loss of precious human lives. Forest fires is one of the major environmental concern that affects the preservation of forests, resulting in economy and ecological damage that causes human suffering. Refer to Elmas [2], quick fire detection and response are the effective way in reducing the damages caused by forest fires, where the various studies have been made in order to improve early fire prediction and detection systems that helps to develop response strategies during the fire. That means, one of key success of putting out forest fires is by providing an early warning detection. Early warning detection is related to accurate prediction of results based on determined parameters. There are three trends technique that could be used in predicting forest fire such as the use of satellite data, infrared or smoke scanners and local sensors, for example, using the meteorological [1].

However the time series data as a knowledge for predicting forest fire is still limited, beside that the data is provide by offline access that actually accessed by downloading the bulk of data. Today with the advancement web of data technology, we could gathering an online data, moreover in real-time or live data, where the data sources is accountable because it comes from the government. With the available semantically annotated data, the interest in extracting valuable information from this area receives increasing attention. Most classical machine learning approaches they require certain attributes that have nominal or numerical values assigned as source of dataset. In contrast, semantic data is based on a RDF graph model which is represent with triple tuples there are subject, predicate, and object. The example Linked Open Data (LOD) could be seen in TWC LOGD [3] that provide an open access to government data or also known Linked Open Government Data (LOGD).

'Linked Data' was defined by Tim Berners-Lee [4], as a number of publish RDF graphs that can be navigated across servers by following the links in the graph in a manner similar to the way the HTML Web is navigated. Refer to Ding [5], Data.gov is a website that provides US Government data to the general public that ensure better accountability and transparency. Ding et al, was build Data-gov Wiki, which attempts to integrate the datasets published at Data.gov into the Linking Open Data (LOD) cloud that have 5 billion triples, where in his work, until they build TWC LOGD [3], as a portal to access any linked Open Government Data from Data.gov.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Based on Paulheim [6], Linked Open Data could be used as background knowledge in data mining. Where the number of feature that relate with the objective of data mining can be retrieve. There are two principle strategies for using Linked Open Data for data mining: (1) Developing specialized mining methods for Linked Open Data. (2) Pre-processing Linked Open Data so that it can be accessed with traditional data mining methods.

Linked Open Data can be related as source of online knowledge in data mining, especially in predicting many real world problems or to test any scientific understanding of the behavior of complex systems or phenomena. The predictions could be used as a guide or basis for decision making [1]. According to Iliadis [7] based on the perspectives of forest fires, several scientists around the world had performed statistical approaches such as regression analysis, probabilistic analysis and artificial intelligence approaches in predicting the forest fire. Some data mining techniques have been applied in the domain of fire detection, for example adopted meteorological data to predict forest fire [1]. Back propagation neural network and the rule generation approach [8], fuzzy c-means clustering application in the case of forest fires [7], artificial neural network to the real word problem of predicting forest fires [9], Neural Network (NN) and Support Vector Machines to predict forest fire occurrence based on weather data [10], classifying hotspot location using Decision Tree Algorithm [11]. However, there is no any research that concern in using Linked Open Government Data, as Background Knowledge that provide data for predicting Forest Fire.

In this research, we have motivation in utilizing Linked Open Government Data (LOGD) as Background Knowledge that provide time series data for prediction Number of Forest Fire and Burn Area Size of Forest Fire. This research describe overall process in accessing and extracting LOGD with SPARQL Endpoint into Row and Colom time series data, then machine learning such as Linear Regression, Back-Propagation Neural Network and Support Vector Machine will be used to provide online prediction.

The reminder of this page could be seen as follow: chapter two talking about related work, chapter three talking fundamentals, chapter four describing the overall proposed method we used to predict number *of forest fire* and *size of Burn Area* from Linked Open Government Data, chapter five describing the result experiment and discussion, the last chapter is conclusion and future work of this research.

2 RELATED WORK

Refer to [12] the used of Semantic Web (LOGD) and data mining itself has emerged as sub field of Semantic Web Mining that focus in mining the documents in web of data. There are several research that works with linked open data in data mining such as Mencia et al [13], they introduce a general approach in order to convert linked data in a relational format which can be used by traditional machine learning approaches, where the results will give complex procurement information as a support strategic decision making based on World Wide Web information.

Heiko Paulheim [14] has present a fully automatic approach for enriching data with features that are derived from Linked Open Data that also produce open-source tool FeGeLOD. They give simple feature generation technique that has been used in his study, the generated features may help improve the results in data mining tasks, where his approach also can be used in the fields of ontology learning and ontology matching.

Mathieu d'Aquin [15] has present an initial approach for exploring open Linked Data sources in interpreting the results of a data mining method, as part of a Learning Analytics process. They demonstrate on a use case relying on data about students' enrolment in course modules how results from a sequential pattern mining process can be automatically organized in a variety of dimensions, obtained from a connected Linked Data source. They also discuss the advantages his approach that combining data mining and Linked Data for Learning Analytics.

Faninzi [16] in his research, they present learning machine approach to mining some information in Linked Open Data through self-training strategy. Several algorithm such as C 4.5, K-NN, JRip, NB, SVM has been applied in his study to find out the best algorithm for mining Linked Open Data through self-training strategy. Moreover Tsouklas et al [17] in his research proposed API that can be used for extracting implicit information from linked dataset Web of Data based on the usage of some network analysis algorithms on a Linked Dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

There are several research that focus in predicting forest fire with machine learning such as conduct by Satoh et al [18], developed a system for predicting the dangers of forest fires. A simulation of dangers related to forest fires was created not only using the weather conditions before, but also coupled with data on population density and so on. Cortez in his research [1], use of five different data mining techniques to predict the burned area of forest fires using Support Vector Machines (SVM) and Random Forests. With the four distinct features likes spatial, temporal, Fire Weather Index components and weather variable such as temperature, relative humidity, rain and wind it is found that the best configuration uses a Support Vector Machine, which is capable of predicting the frequent burned areas due to small fires.

Iliadis et al [7], they conducted a study to increase the fuzzy c-means model intelligently using a flexible termination criteria. For the clustering of forest fires. This way enables the algorithm to be more flexible and human like in an intelligent way. It also avoids possible infinite loops and unnecessary iterations. Sitanggung et al [11], applied decision tree C4.5 algorithm to predict the location of the incident hotspots in Rokan Hilir district, Riau province, Indonesia. The dataset consists of hotspot occurrence locations, human activity factors, and land cover types. Human activity factors include city center locations, road network and rivers network. The results indicated a decision tree which contains 18 leaves and 26 nodes with an accuracy of about 63.17%.

Safi et al in his research [9], applying artificial neural networks to the real world problem of predicting forest fires, using the back propagation learning algorithm. The synaptic weights of this architecture were adjusted with back propagation algorithm. Yu et al [8], conducted a research investigating the nonlinear relationship between the size of a forest fire and meteorological variables (temperature, relative humidity, wind speed and rainfall) using two hybrid approaches namely the historical meteorological variables clustering using Self Organizing Map. Back-propagation neural network and the rule generation approaches to get a different input from the clustered data.

Sakr et al in his first publication [19] applied a description and analysis of forest fire prediction methods based on artificial intelligence. A novel forest fire risk prediction algorithm, was presented based on Support Vector Machines to predict the fire hazard level of a day, where the algorithm depended on previous weather conditions. In his second publication [10], they reducing a set of weather parameters that utilizing relative humidity and cumulative precipitation to estimate the risk of the output, to predict the occurrence of forest fires by comparing two artificial intelligence based methods that is the Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

Based on the information above, there has been no any research that predict number of forest fire and Burn Area with the sources of data based on Linked Open Government Data (LOGD). Since the technology web of data also categorized as new field, there has been no any research that concern in utilize LOGD as background knowledge especially in predict forest fire. With Linked Open Government Data, there is possibilities to conducting online data mining.

3 FUNDAMENTAL

3.1 Forest Fire

Forest fires are natural disasters that occur because of a fire that destroyed forests, and can make great danger to people who live in forests as well as to wildlife. Forest fire can occur by lightning, human negligence or arson that can burn thousands of square kilometers. According to Brown and Davis [10], there are three types of forest fires namely: ground fire, surface fire and crown fire.

There is a relationship between meteorological conditions and forest fire in some studies, where meteorological variables, such as temperature, relative humidity, wind speed and precipitation are believed related with forest fire [1]. For the sample, meteorological variables are used to measure the fire index rating that was applied in Canadian forest fire weather index (FWI) based on Temperature, Relative Humidity, Rain, Wind, Fine Fuel Moisture Code, Duff Moisture Code, Drought Code, and Initial Spread Index. This index was adopted and used by several countries [20].

Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

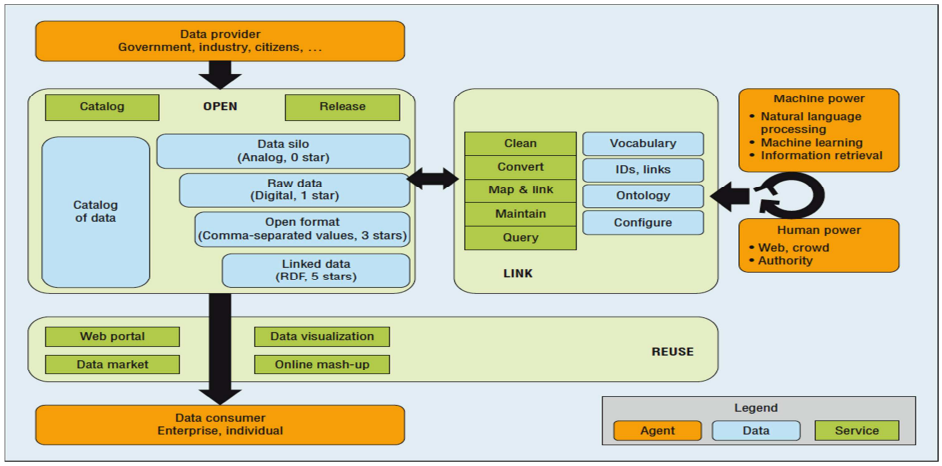

Figure. 1. Concept of linked open government data [21].

3.2 Linked Open Government Data (LOGD)

Linked Open Government Data (LOGD) is the advancement from Linked Open Data (LOD) concept that applied in Government area, where basically is a number of publish RDF graphs that can be navigated across servers by following the links in the graph that similar like navigating HTML Web.

Refer to Ding et al [21] LOGD has been start by Data.gov and Data.gov.uk that linked Web with government data as a way of facilitating opening, linking. LOGD represents a new data integration paradigm for sustainable growth of OGD that allowing users to mash up government data with crowd sourced data, private owned data, and many other types of non-governmental data. Besides that, since the LOGD adapted data oriented architecture (DOA) in principal anyone can make contribution to LOGD deployment. LOGD can be accessed with SPARQL query from its SPARQL Endpoint.

According to Tim Berners-Lee's, LOGD is recognized as a Web-based open ecosystem that organically interconnects the original data owners (e.g. government agencies), data-processing service providers (e.g. entity resolution services), and data consumers (e.g. enterprises and citizens). Figure 1 shows the roadmap concept of LOGD with three data-processing stages there are [21]:

- Open stage, where government agencies play a key role in putting OGD datasets online in reusable formats and maintaining central OGD catalogs to help citizens finding available and relevant datasets
- Link stage, community participants (industry and academia, for example) help enhance the quality of the released OGD data.
- Reuse stage, developers pull the published OGD datasets together to build high-value applications.

3.3 SPARQL Query

SPARQL can be used to express requests across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and questions constrained by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs [22].

Refer to Sirin [23], the vocabulary for RDF graphs is basic three disjoint sets there are: a set of URIs V_{uri}, a set of bnode identifiers V_{bnode}, and a set of well-formed literals V_{lit} uri. The union of these sets is called the set of RDF terms. An RDF have triple tuple $(s,p,o) \in (V_{uri} \cup V_{bnode}) \times V_{uri} (V_{uri} \cup V_{lit})$. An RDF graph is a finite set of RDF triples, where the over all building block for SPARQL queries is Basic Graph Patterns (BGP). More complex SPARQL queries are constructed from BGPs by using projection (SELECT operator), left join (OPTIONAL operator), union (UNION operator) and constraints (FILTER operator). The detail explanation of SPARQL Query could be seen at [22].

Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The uses of SPARQL Query in this research is for extracting and building background knowledge to become Row and Colom time series data with its attribute, where the process of extracting by accessing the SPARQL Endpoint from SPARQL Iterator.

3.4 Data Mining

3.4.1 Preprocessing

There are several process in preprocessing data refer to Han [24] such as data cleansing, data integration, data reduction and data transformation. In this research we doing data transformation that will be used to normalize the data. In this research we used Min-max normalization that performs a linear transformation on the original data. The time series data will be normalize with the max value is 1 and the minimum value is 0. Suppose that minA and maxA are the minimum and maximum values of an attribute, A. Min - max normalization maps a value, v_i, of A to $v'i$ in the range $[new\_min\ A, new\_max]$ where the formula could be seen at (14).

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(New\_max_A - New\_min_A) + New\_min_A \quad (1)$$

3.4.2 Time Series Data and Window Size.

A time series is a sequences of vector (t), $t = 0,1,\ldots.$, where t represent elapsed time, x will be sampled to give a series of discrete data points, equally spaced in time. The size of the time interval usually depends on the problem that can be anything in time format from milliseconds, hours to days, or even years. The sampling frequency in time series have important roles, because different frequencies can essentially change the main characteristics of the resulting time series. In many cases, the observer sampling time series data could be at discrete time, accumulated time interval or averaged time interval.

Window size in time series data is process to transform the data to have more attribute in time series manner as showed in Figure 2. Refer to frank [25], if the window size is too small then the attractor of the system is being projected onto a space of insufficient dimension, in which proximity is not a reliable guide to actual proximity on the original attractor. Moreover, a window of too large a size may also produce problems: since all necessary information is populated in a subset of the window, the remaining fields will represent noise or contamination.

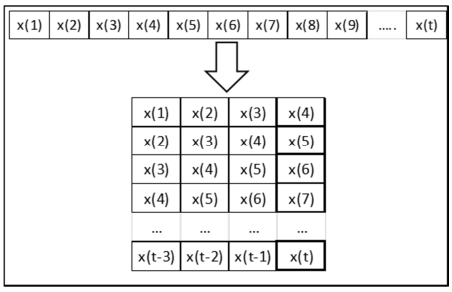

Figure. 2. Example Of Window Size Time Series Data.

3.4.3 Prediction Algorithm.

According to Witten et al [26], data mining is defined as the process of discovering patterns in large volume of data that have meaningful information, where the process must be automatic or semiautomatic. There are several algorithm that capable in predicting or forecasting time series data, such as Neural Network, Linear Regression, and Support Vector Machine SVM. In this research we used those algorithm to find out which algorithm that have best prediction results. Where the detail information about those algorithm could be showed below.

3.4.3.1 Linear regression model

Linear regression model is a statistical technique that has been applied in various fields. The model is one of the most commonly used methods for forecasting [27], [28] the regression model describes the mean of the normally distributed dependent variable y as a function of the predictor or independent variable x [27]:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (2)$$

where y_t is the value of the response or dependent variable from the its pair, b_0 and $b_1$are the two unknown parameters, x is the value of the independent variable from the tth pair, and ε_t is a random error term. The estimated values of the regression model are calculated in (3), where b_0 is intercept and b_1 is slope parameter:

$$y_t = b_0 + b_1 x_t \quad (3)$$

3.4.3.2 Neural network model

Neural Networks have been widely used as time series prediction: which employ a sliding window over the input sequence. The examples of this ap-

Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

proach has been applied such as in market predictions, meteorological and network traffic forecasting. Back-Propagation Neural Network BPNN is a Neural Net forecasting model that capable in handling nonlinear, where the output value of a multilayer back propagation neural network is computed as [28]:

$$Y(x) = \beta_0 + \sum_{J=1}^{H} \beta_j \psi\left(\gamma_{j0} + \sum_{i=1}^{n} \gamma_{ji}\, x_i\right) \quad (4)$$

Where $(\beta_0, \beta_1, \ldots . \beta_H,\ \gamma_{10}, \ldots . \gamma_{Hn})$ are the weights or parameters of the neural network, and the symbol ψ is the activation function, where in this research we uses sigmoid activation function.

3.4.3.3 Support vector machine

Support vector machine SVM is known as the algorithm that finds a special kind of linear model with the maximum margin hyperplane, where the maximum margin hyperplane gives the maximum separation between the decision classes. The training examples that are closest to the maximum margin hyperplane are called support vectors. All other training examples are irrelevant for defining the binary class boundaries. For the linearly separable case, a hyperplane separating the binary decision classes in the three-attribute case can be represented as the following equation:

Refer to Kim, [29] in linear separable case, a hyperplane for separating the binary decision classes can be represented in equation (5):

$$y = w0 + w1x1 + w2x2 + w3x3 + \ldots\ldots + wixi \quad (5)$$

Where y is the outcome, x_i is the attribute and w_i is the weight parameter that determine the hyperplane. The maximum hyperplane in term support vector could be calculate with equation (6):

$$y = b + \sum \alpha i\, \gamma i\, \chi(i)\, .\, \chi \quad (6)$$

Where b and αi are parameter that determine the hyperplane, than γi is the class of training $\chi(i)$. The parameter χ is representing a test example and the vector $\chi(i)$ are the support vector. For implementing SVM in nonlinear class we can use equation (7) to find out the maximum hyperplane.

$$y = b + \sum \alpha i\, \gamma i\, K(\chi(i).\chi) \quad (7)$$

The function $K(\chi(i).\chi)$ is defined as kernel in SVM for generating inner product to construct machine in nonlinear input. The kernel SVM have several type such as based on polynomial, Gaussian radial function and etc.

4 PROPOSED METHOD

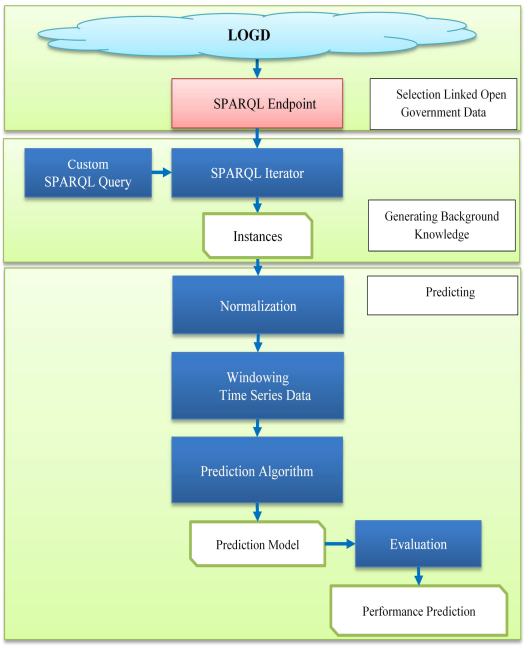

Figure 3. Proposed Method

4.1 Linked Open Government Data

In this phase we selecting which Linked Open Government Data Portal that could be access to provide background knowledge of time series data to predict number of Forest Fire and Burn Area Forest Fire. We find out Tetherless World Constellation Linked Open Government Data (TWC LOGD) [3] as a sources portal that have accountable data from open government United States of America http://data.gov, especially from Department of The Interior.

TWC LOGD has define as LOGD ecosystem as a Linked Data-based system where stakeholders of different sizes and roles find, manage, archive, publish, reuse, integrate, mash-up, and consume open government data in connection with online tools, services and societies as showed in Figure 4.

Using TWC LOGD the valuable open government data that have been recognized by the US government data-sharing community could be provide by Linked Data and Semantic Web technologies. The dataset in TWC LOGD could be access directly from it SPARQL Endpoint using SPARQL

Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

query language. The detail datasets catalogue that relate with forest fire in TWC LOGD could be showed at http://data-gov.tw.rpi.edu/wiki/. We search which RDF (linked data) that contain with time series Forest or Wildland Fire data, than with simple SPARQL query bellow we check its variable. From this portal, we could gather open government time series data of Forest Fire from 1960 until 2008.

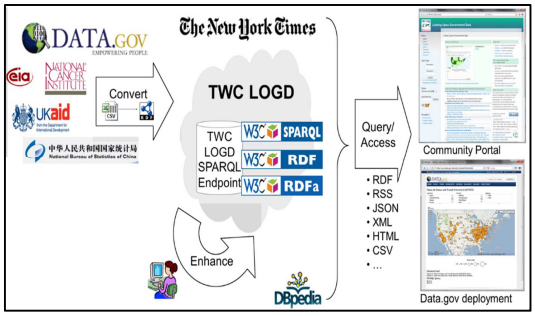

Figure 4. Workflow Of The TWC LOGD Portal [3]

```
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#

SELECT ?attribute ?g
WHERE{
GRAPH http://data-gov.tw.rpi.edu/vocab/Dataset_1187
{
?g rdfs:label ?attribute.
}
}
```

4.2 SPARQL Query and Iterator

In this step we access the SPARQL Endpoint of TWC LOGD with SPARQL Query. There are several tools that has been built by existing research as Iterator to access The SPARQL Endpoint such as in [14], [17]. RapidMiner Linked Open Data Extension[1] is used as SPARQL Iterator that capable interfacing selected endpoint and delivery the query results into Row and Colom instance.

```
PREFIX data: http://data-gov.tw.rpi.edu/vocab/p/1187/

SELECT ?year ?fires
WHERE{
 GRAPH <http://data-gov.tw.rpi.edu/vocab/Dataset_1187>
  {
 ?o data:year ?year .
 ?o data:fires ?fires .
  }
}
```

```
PREFIX data: http://data-gov.tw.rpi.edu/vocab/p/1187/

SELECT ?year ?burn_area
WHERE{
 GRAPH <http://data-gov.tw.rpi.edu/vocab/Dataset_1187>
  {
 ?o data:year ?year .
 ?o data:acres ?burn_area .
  }
}
```

The focus in this phase is accessing two type data, the first is Number of Forest Fire data and the second is Burn Area Size data. The detail SPARQL query that used in building Row and Colom data are showed in the upper part. The results of this phase are the instance data that consist with attribute and time series data, where the instant could be showed in Figure 5.

year	fires
1960	103387
1961	98517
1962	115345
1963	164183
1964	116358
1965	113684

(a)

year	burn_area
1960	4478188
1961	3036219
1962	4078894
1963	7120768
1964	4197309
1965	2652112

(b)

Figure 5. Results of Instance SPARQL Query, (a) time series data of Number Forest Fire, (b) time series data of Burn Area Size

4.3 Normalization

After we retrieve the results from SPARQL Query instant, we continue with the process of normalization. Normalization process in this research using formula (1) with min max normalization.

The normalization process only transform '*fires*' or '*burn_area*' attribute, the attribute '*year*' have not normalized because it used as (*id*). These process will giving results minimum and maximum value between [0, 1] in its time series data. The sample of process before and after normalization could be seen in Figure.3 and Figure.4.

[1]http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/

Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

year	fires
1960	0.368
1961	0.347
1962	0.420
1963	0.631
1964	0.425
1965	0.413

(a)

year	burn_area
1960	0.382
1961	0.216
1962	0.336
1963	0.684
1964	0.349
1965	0.172

(b)

Figure 6. Results Instances After Normalization, (A) Time Series Data Of Number Forest Fire, (B) Time Series Data Of Burn Area Size

4.4 Window size of Time Series Data

In this part we transform the number fire and burn area time series data in several window slide to produce more attribute into four types size of window. We make the size of window into 2, 4, 8, and 10 window size. The sample of results in process window size could be showed in Figure 7.

year	fires-1	fires-0
1960	0.208	0.368
1961	0.552	0.347
1962	0.462	0.420
1963	0.505	0.631
1964	0.390	0.425
1965	0.460	0.413
1966	0.236	0.451

(a)

year	fires-3	fires-2	fires-1	fires-0
1961	0.420	0.338	0.552	0.347
1962	0.236	0.451	0.462	0.420
1963	0.278	0.431	0.505	0.631
1964	0.505	0.631	0.390	0.425
1965	0.299	0.272	0.460	0.413
1966	0.292	0.263	0.236	0.451
1967	0.263	0.236	0.451	0.462

(b)

Figure 7. (a) Window Size n=2 (b) Window Size n=4

4.5 Prediction Algorithm

After the process of windowing time series data that has been extract from LOGD, the instances data is ready to proceed by machine learning. Three machine learning such as Back Propagation Neural Network BPNN, Linear Regression and Support Vector Machine SVM are applied in this phase. The detail method explanation of each algorithm could be showed in fundamental. The experiment is done with RAPIDMINER[2], we setup the parameter of Linear Regression, BPNN and SVM to perform satisfaction results in predicting Number Forest Fire and Burn Area:

- Linear Regression :
  - M5-prime is set as Feature Selection during regression,
  - Minimum tolerance of eliminate collinear feature is set as 0.05.

- Back propagation neural network Parameter :
  - The momentum size is set as 0.3,
  - Learning rate as 0.001,
  - Training cycles size as 1000.

- Support Vector Machine Parameter :
  - The Radial Kernel is set as kernel,
  - Kernel Gama is set as 7.0,
  - Kernel Cache is set as 200,
  - Convergence epsilon 0. 1,
  - Maximum iteration is set as 100000.

The proposed method in predicting Number of Forest Fire and Burn Area in this research, can be applied to others prediction time series field research that using Linked Open Government Data as Background Knowledge for data sources.

5 RESULTS & DISCUSSION

5.1 Performance Measures

The prediction or forecasting models are evaluated in terms of their ability to forecast the future values. The performance of prediction algorithm in this research using mean square error (MSE) and root mean square error (RMSE) as standard that calculate the number of error rate prediction, where the equation could be showed below:

$$MSE = \frac{\sum_{t=1}^{n}(y_t - y'_t)^2}{n-p} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - y'_t)^2}{n-p}} \quad (9)$$

5.2 Results Experiment

In this research, based on the graphical representation used in context of the RDF data model in LOGD, prediction time series model with machine

[2] http://rapidminer.com/products/rapidminer-studio

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

learning techniques are applicable for predicting Forest Fire. Result experiment in this study has capable in using linked open government data LOGD as background Knowledge in provide time series forest fire data, where the data consist with number forest fire and burn area. LOGD was given an easy ways in accessing accountable open government data from its SPARQL Endpoint.

The overall prediction performance of number forest fire and burn area size forest fire could be showed in Table 1 and Table 2. The results comparison of three prediction model have showed that if Linear Regression algorithm capable to achieve best prediction results of Number Forest Fire using 2 window size with MSE 0.065 and RMSE 0.255. In the other hands, Support Vector Machine SVM capable in resulting best prediction performance in predicting Burn Area of Forest Fires using 4 window size with MSE 0.043 and RMSE 0.207. The performance of Back-Propagation Neural network in this study unable to achieve the best result, but this method is stable in predicting both data that showed its performance always have second position compare with other algorithm.

Table 1. Predicion Performance Of Number Forest Fire

ALGORITHM	Window Size	RMSE	MSE
Linear Regression	***2***	***0.255***	***0.065***
Linear Regression	4	0.274	0.075
Linear Regression	8	0.797	0.636
Linear Regression	10	0.586	0.343
BPNN	2	0.266	0.071
BPNN	4	0.277	0.077
BPNN	8	0.276	0.076
BPNN	10	0.378	0.143
SVM	2	0.311	0.097
SVM	4	0.316	0.100
SVM	8	0.297	0.088
SVM	10	0.298	0.089

Table 2. Prediction Performance Of Burn Area

ALGORITHM	Window Size	RMSE	MSE
Linear Regression	2	0.272	0.074
Linear Regression	4	0.210	0.044
Linear Regression	8	0.303	0.092
Linear Regression	10	0.300	0.090
BPNN	2	0.263	0.069
BPNN	4	0.210	0.044
BPNN	8	0.277	0.077
BPNN	10	0.253	0.064
SVM	2	0.290	0.084
SVM	***4***	***0.207***	***0.043***
SVM	8	0.276	0.076
SVM	10	0.243	0.059

Figure 8 and Figure 9 are showed the gap between Actual data and the prediction results of number forest fire and burn area forest fire prediction. In Figure 8 is showed Linear Regression that used two window size to predict number of forest fire. Moreover, in Figure 9 is showed the prediction of Support Vector Machine in predict size of burn area forest fire data, where the gap is not too far with the actual data. The result in this study have showed if Linked Open Government Data has able to be used as Background Knowledge to provide accountable Forest Fire data from the government, where the results in this study also give promising result in predicting Forest Fire.

6 CONCLUSION AND FUTURE WORK

Linked Open Government Data in this study has able to be used as background knowledge in providing time series data especially number of forest fire and size of burn area forest fire. This research has showed promising results in predict both data.

The best prediction performance in predict number of forest fire data has achieved by Linear Regression that showed MSE 0.065 and RMSE 0.255, more over Support Vector Machine SVM was able to perform the best prediction with MSE 0.043 and RMSE 0.207 in burn area size data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The method that used in this research, is applicable to others time series prediction field research that used Linked Open Government Data as Background Knowledge for data sources. For the future work, optimization technique could be considered to improve the prediction performance, besides that linked open sensor that provide live data will be used for real time prediction.

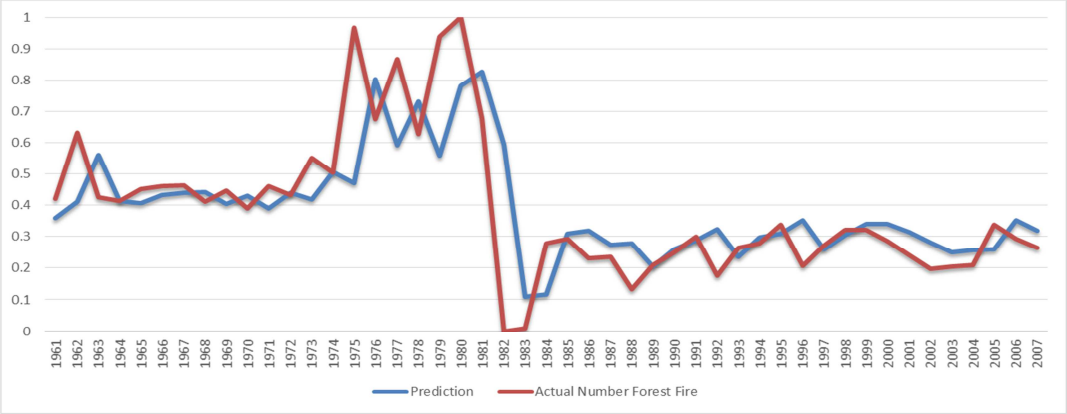

Figure 8. Number Of Forest Fire Prediction With Linear Regression In Window Size =2

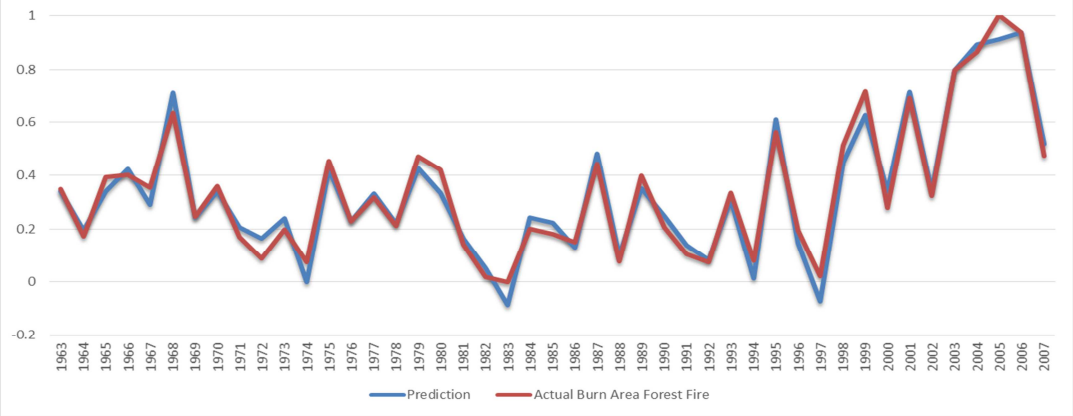

Figure 9. Burn Area Forest Fire Prediction With SVM In Window Size =4

REFERENCES:

[1] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," in *Procedings of 13 Portugese Conference on Artificial Intelligence*, 2007.

[2] C. Elmas and Y. Sonmez, "A data fusion framework with novel hybrid algorithm for multi-agent Decision Support System for Forest Fire," *Expert Systems with Applications journal,* vol. 38, p. 9225–9236 Contents, 2011.

[3] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Wil-liams, X. Li, J. Michaelis, A. Graves, J. G. Zheng and Z. S. a. J. F. a. D. L. M. a. J. A. Hendler, "TWC LOGD: A portal for linked open government data ecosystems," *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 9, no. 3, pp. 325-333, 2011.

[4] T. Berners-Lee, "Design Issues: Linked Data," W3C, 18 06 2009. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed 1 12 2013].

[5] L. Ding, D. DiFranzo, A. Graves, JamesMichaelis, X. Li, D. L. McGuinness and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

J. Hendler, "Data-gov Wiki: Towards Linking Government Data.," in *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, 2010.

[6] H. Paulheim, "Exploiting Linked Open Data as Background Knowledge in Data Mining," 2013.

[7] L. Iliadis, M. Vangeloudh and S. Spartalis, "An intelligent system employing an enhanced fuzzy c-means clustering model: Application in the case of forest fires.," *Computers and Electronics in Agriculture,* vol. 70, no. 2, pp. 276-284, 2010.

[8] Y. P. Yu, R. Omar, R. D. Harrison, M. K. Sammathuria and A. R. Nik, " Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods," *Journal of Computational Biology and Bioinformatics Research,* vol. vol. 3, pp. 47-52, 2011.

[9] Y. Safi and A. Bouroumi, "A neural network approach for predicting forest fires.," in *IEEE Multimedia Computing and System*, 2010.

[10] G. E. Sakr, I. H. Elhajj and G. Mitri, "Efficient forest fire occurrence prediction for developing countries using two weather parameters," *Engineering Applications of Artificial Intelligence,* vol. vol 24, no. 5, p. 888–894, 2011.

[11] I. S. Sitanggang and M. H. Ismail, "Hotspot occurrences classification using decision tree method: Case study in the Rokan Hilir, Riau Province, Indonesia," in *International Conference in ICT and Knowladge Engineering IEEE* , Bangkok, 2010.

[12] B. Berendt, A. Hotho and G. Stumme, "Towards semantic web mining," in *in Proceedings of the 1st International Semantic Web Conference, ISWC2002*, 2002.

[13] E. L. Mencia, S. Holthausen, A. Schulz and F. Janssen, "Using Data Mining on Linked Open Data for Analyzing E-Procurement Information - A Machine Learning approach to the Linked Data Mining Challenge 2013," in *CEUR Workshop Proceedings*, 2013.

[14] H. Paulheim and J. mkranz, "Unsupervised Generation of Data Mining Features from Linked Open Data," in *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics - ACM*, New York, 2012.

[15] N. J. Mathieud'Aquin, "Interpreting data mining results with linked data for learning analytics: motivation, case study and directions," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge - ACM*, 2013.

[16] N. Fanizzi, C. dAmato and F. Esposito, "Mining Linked Open Data through Semi-supervised Learning Methods Based on Self-Training," in *IEEE Sixth International Conference on Semantic Computing (ICSC)*, 2012.

[17] C. Tsoukalas, D. Dervos, J. Martinez-Gil and J. F. Aldana-Montes, "TheMa: An API for Mining Linked Datasets," in *16th Panhellenic Conference on Informatics (PCI)*, 2012.

[18] K. Satoh, W. Song and K. T. Yang, "A Study of Forest Fire Danger Prediction," in *15th International Workshop on Database and Expert Systems Applications IEEE*, 2004.

[19] G. E. Sakr, I. Elhaj and G. Mitri, "Artificial intelligence for forest fire prediction," in *IEEE in Advanced Intelligent Mechatronics*, 2010.

[20] W. Groot, R. Field, M. Brady, O. Roswintiarti and M. Mohamad, "Development of Indonesia and Malaysia Fire Danger Rating System," *Mitigation and Adaptation Strategies for Global Change, Springer Link,* vol. 12, no. 1, pp. 165-180, 2007.

[21] L. Ding, V. Peristeras and M. Hausenblas, "Linked Open Government Data," *IEEE Intelligent Systems,* vol. 27, no. 3, pp. 11-15, 2012.

[22] E. Prud'Hommeaux and A. Seaborne, "SPARQL query language for RDF," *W3C recommendation,* vol. 15, 2008.

[23] E. Sirin and B. Parsia, "SPARQL-DL: SPARQL Query for OWL-DL," in *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*, Insbruck, Austria, 2007.

[24] J. Han, M. Kamber and J. Pei, DATA MINING: CONCEPTS AND TECHNIQUES 3RD EDITION, Morgan Kaufman, 2011.

[25] R. J. Frank, N. N Davey and S. P. Hunt, "Input window size and neural network predictors," in *IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN*, 2000.

[26] I. H. Witten, E. Frank and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques, Elsevier, 2011.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[27] C. T. Le, Introductory Biostatistics, New Jersy: John Wiley and Son, 2003.

[28] Purwanto, C. Eswaran and R. Logeswaran, "A dual hybrid forecasting model for support of decision making in healthcare management," *Advances in Engineering Software,* vol. 53, no. 0, pp. 23 - 32, 2012.

[29] K.-j. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing,* vol. 55, pp. 307 - 319, 2003.