**CHAPTER 2**

This chapter explain about the literature review of this project from different of authors view from different studies .The methods that are being explained in this chapter that are consist of project domain such as web mining, mobile application and many others method that involved in this project.

**2.1 DATA MINING**

Data mining is a method to analyze and accumulating the data from different perspective by extraction of knowledge from huge amount of data in order to get useful information .data mining extract the information from existing data and present the new useful data or information hat help to do the decision making (Han, Pei, & Kamber, 2011; Kuşak, 2014; Wasilewska & Menasalvas, 2018). The process of data mining is include the process of creating target data or specify the data target, data preprocessing, mining proper, evaluation of the pattern and present the new knowledge. Mathematical modelling and algorithm used in mining the data in order to segment the data and predict the values through searching the large collection of data to detect pattern or trends(Mauricio, Payawal, Cueva, & Quevedo, 2016). Data mining also used in weather prediction it predict the atmosphere condition in specific region at an unspecified time in future. World Wide Web and other advanced database systems, and it can be used to do analysis (Mazhar, Ikram, Butt, & Butt, 2015).

**2.1.1 TEXT PROCESSING**

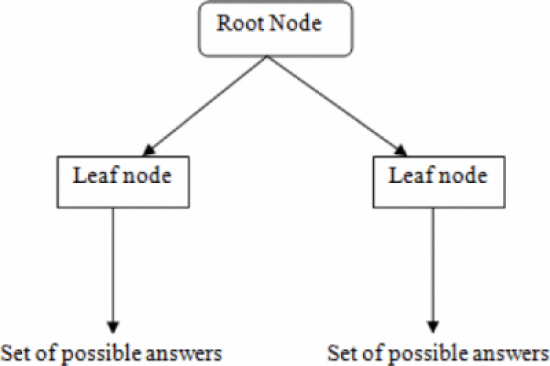
Text processing is a part of natural language processing and data mining. Text processing is the most suitable way to extract data and information from specific source such as text file and html document(Nelson, 2017) .with the growing of internet that provide huge amount of text information that lead the popularity of the text processing. majority of the application of text processing are include the process of text searching and text replacement .text processing usually use regular expression to find the specific text and extract the substring from the data source such as text file (Song, Huang, & Liang, 2009).Text extraction has many usage like sentiment analysis and fact extraction. In order to determine the significance of the word in the document and counting on how many the term appear in the document is the text processing process for example the use of Term frequency-Inverse Document Frequency (TF-IDF).text processing is very useful in the area of article analysis and image retrieval by using image caption (Ramisa, Yan, Moreno-Noguer, & Mikolajczyk, 2018)

**2.1.2 CLASSIFICATION**

Classification is a data analysis technique that extract model in order to describe the important data classes. The classifier predict categorical class labels whether it’s discrete or unordered. The model or classifier are constructed during the task analysis phase in order to predict the class label attribute or categorical label. Classification is aim to find the relationship between the set of descriptive variable and the categorical response. Prediction of the membership of unknown samples to the defined category need identification of relationship between variable and a class vector. Classification of data allows a better understanding of unstructured data as it predicts the data’s categorical labels. Classification is a supervised learning data analysis method, which means the predictive model learns from a training set that has its label that is discrete-valued and unordered, accompanied with the data, indicating the class (Ballabio, Grisoni, & Todeschini, 2018; Han, Kamber, & Pei, 2012; Kesavaraj & Sukumaran, 2013)

**2.1.2.1 DECISION TREE**

Decision tree is the technique used in classification especially in multi-category classification it is a tree-like structure that contain root node and branch that represent the result of the test, leaf hold the class label(Wang, Wang, Nie, Yu, & Wang, 2018). Decision tree classifiers are easy to use because it is easy to convert it into classification rules. The learning and classification steps are fairly simple and fast. Decision trees have been tested in mapping of landslide susceptibility and have resulted in a viability of the method to be used in such field (Tehrany, Pradhan, & Jebur, 2014). The goal of the decision tree is to create a model that able to predict the value of a target variable based on the input. Figure 2.1 illustrate that each of the internal node are corresponding to the input variable and the leaf of the tree represent the target variable (Kesavaraj & Sukumaran, 2013).



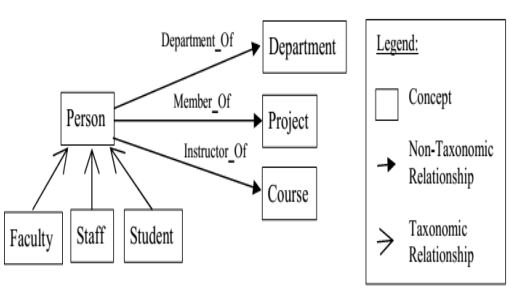
**Figure 2.1** Decision tree induction (Source : Kesavaraj et al.,2013)

**2.1.2.2 SUPPORT VECTOR MACHINE (SVM)**

Support vector machine is the tools used in classification technique and it’s a supervised machine learning algorithm that act as the prediction tools to define classes out of datasets by drawing a decision boundary to easily separate the data. The typical problem that using the support vector machine is the quadratic optimization which affected by dimension of features such as the size of the dataset and the present of the dataset (Somvanshi & Chavan, 2016). SVM is used in both linear and nonlinear problems. For linear problem, SVM draws a decision boundary or hyperplane to differentiate the classes of a dataset, while for a nonlinear problem, kernel function is used to map the data, which requires a large feature space. SVM was used in a project that maps flood susceptibility (Tehrany et al., 2014).

**2.1.2.3 ONTOLOGY**

The term, ontology is related to the philosophical meaning of “existence”, where anything that exists can be represented in human-readable form, it describe the meaning of the names of entities. Ontology allows better understanding of data that supports the integration of data. It has been widely used in knowledge-based system to represent knowledge. An example of it is from a review paper on ontology matching with background knowledge for Linked Open Data (LOD).ontology is the center of knowledge of any domain, it represent an area of knowledge. Ontology consist of attribute, class and relationship that illustrated on Figure 2.2 (Asikri et al., 2017; Husein, Akbar, Sitohang, & Azizah, 2016).



**Figure 2.2** University structure based on ontology representation

**2.1.3 CLUSTERING**

Clustering is the technique under the unsupervised learning and the goal of clustering is to partition the collection of object that have the same similarity among the object into a group that called cluster (Cornuéjols, Wemmert, Gançarski, & Bennani, 2018; Saxena et al., 2017). Distance between clusters are calculated using a dissimilarity function that distinguishes the clusters apart. This way, it is less costly to produce a training set with labels like in classification method. Clustering has been found used in multiple projects like determining effects of clustering techniques in perception of analyst about the risk of portfolios designed for financial stability monitoring (Lemieux, Rahmdel, Walker, Wong, & Flood, 2014).

Other than that, clustering has been incorporated in grouping of students according to their academics, punctuality, and exam in helping to evaluate the students’ performance prediction (Ahmed & Elaraby, 2014). Another study suggested in using clustering techniques in distinguishing the “highs” and “lows” of rainfall in order to overcome constraint urban sewage system (Löwe, Madsen, & McSharry, 2016). Clustering is also used in job offers categorization (Poch, Bel, Espeja, & Navio, 2014).

**2.1.4 WEB SCRAPING**

Web scraping or web data extraction can be done by scraping data from websites. Web scraping is quick, efficient and automated which presents data in a structured and easy-to-use format (Castrillo-Fernández, 2015). The process of scraping data from the web can be accomplished by writing in various programming languages such as Java, Python, and Ruby, or even by using scraping tools also called Web Data Extractor or Crawling Tools. Pentaho, a web scraping software, have been used to extract data from computer shops in objective of recommending computer parts in accordance to people’s needs (Julian & Natalia, 2015).

**2.2 WEB SCRAPING APPROACH**

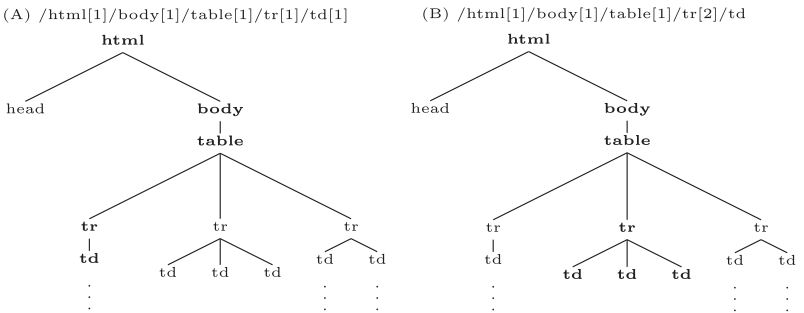
There are several technique under web scrapping method such as tree-based technique, web wrapper approach and hybrid technique(Ferrara, De Meo, Fiumara, & Baumgartner, 2014; Foley, Bendersky, & Josifovski, 2015).

**2.2.1 TREE-BASED TECHNIQUE**

Document Object Model (DOM) can be defined as the representation of web page via well-arranged rooted tree .Usually ,html document is in plain text that contain the tag that can be interpret by the web browser such as Google browser and Firefox. The hierarchy of the tag or label inside the html document captured in Document object model. The tag inside html tag is known as node in Document Object Model tree (Ferrara et al., 2014; Li, Yang, Xiu, & Liu, 2016). There are several technique in tree-based technique such as using Xpath expression, Tree Edit Distance Matching algorithm.

**2.2.1.1 Xpath expression approach**

Xpath is mainly designed for xml document and the purpose of Xpath can be used to navigate the element inside xml document or any document that can be converted into tree that known as Document Object Model (DOM).Xpath can identify single or multiple element inside the DOM tree. Extraction of data from web page is much more easier by using Xpath expression but the problem is when the web page is constantly change from time to time because Xpath expression is related to the structure of the element inside the document (Ferrara et al., 2014).

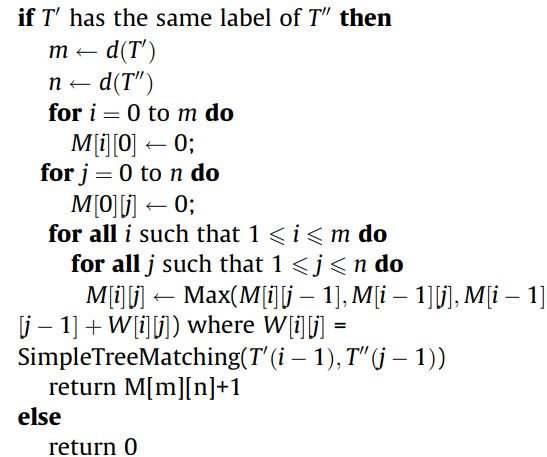
****

**Figure 2.3** example of selecting one node (A) and multiple node (B) in Document Object Model tree

(Source: Ferrara et al. (2014))

**2.2.1.2 Tree Edit Distance Matching**

Tree edit distance matching is an algorithm that designed to calculate the similarity of tree but it require huge amount of memory consumption .Three edit distance compare 2 tree and compute the cost to construct each tree. Simple tree matching is used to solve the problem in tree edit distance matching algorithm but there are limitation of the simple tree matching algorithm such as crossing level is not allowed and cannot match permutation of the nodes (Ferrara et al., 2014; Pawlik & Augsten, 2016).

****

**Figure 2.4** Simple tree matching algorithm used in tree edit distance matching

(Source : Ferrara et al. (2014))

**2.2.2 WEB WRAPPER APPROACH**

Web wrapper approach is a program that provide structure and semi-structure view and it’s implemented in many algorithm that seek of information and data extraction. There are several technique in web wrapper such as regular expression-based approach (Ferrara et al., 2014).

**2.2.2.1 REGULAR EXPRESSION-BASED APPROACH**

Regular expression is a method for string searching and modification of string such as character replacement inside the string. Regular expression can be used to parse file, validate the user input and also access control (Chapman, Wang, & Stolee, 2017). The adoption of regular expression in web wrapper approach have the limitation such as the lack of flexibility for example when the structure of the web page have been change , the rule of regular expression to extract data must be modified(Ferrara et al., 2014).

**2.2.3 HYBRID APPROACH**

Hybrid approach combination of wrapper induction and wrapper generation provided by Roadrunner.it will automatically generate the template for data extraction. Roadrunner work with the data that given by the user and the page that have been generated will automatically generated. Roadrunner exploit the know priority schema on web pages to build knowledge. web data extraction also can be done by using spatial reasoning that use visual model to extract the data (Ferrara et al., 2014; Kurte, Durbha, King, Younan, & Vatsavai, 2017).

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology, 2*(2), 43-47.

Asikri, M. E., Krit, S., Chaib, H., Kabrane, M., Ouadani, H., Karimi, K., . . . Elbousty, H. (2017, 8-10 May 2017). *Mining the Web for learning ontologies: State of art and critical review.* Paper presented at the 2017 International Conference on Engineering & MIS (ICEMIS).

Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems, 174*, 33-44. doi:<https://doi.org/10.1016/j.chemolab.2017.12.004>

Castrillo-Fernández, O. (2015). Web Scraping: Applications and Tools.

Chapman, C., Wang, P., & Stolee, K. T. (2017, Oct. 30 2017-Nov. 3 2017). *Exploring regular expression comprehension.* Paper presented at the 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE).

Cornuéjols, A., Wemmert, C., Gançarski, P., & Bennani, Y. (2018). Collaborative clustering: Why, when, what and how. *Information Fusion, 39*, 81-95. doi:<https://doi.org/10.1016/j.inffus.2017.04.008>

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems, 70*, 301-323. doi:10.1016/j.knosys.2014.07.007

Foley, J., Bendersky, M., & Josifovski, V. (2015). Learning to Extract Local Events from the Web. 423-432. doi:10.1145/2766462.2767739

Han, J., Kamber, M., & Pei, J. (2012). 8 - Classification: Basic Concepts *Data Mining (Third Edition)* (pp. 327-391). Boston: Morgan Kaufmann.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Husein, I. G., Akbar, S., Sitohang, B., & Azizah, F. N. (2016, 26-27 Oct. 2016). *Review of ontology matching with background knowledge.* Paper presented at the 2016 International Conference on Data and Software Engineering (ICoDSE).

Julian, L. R., & Natalia, F. (2015, 25-27 Nov. 2015). *The use of web scraping in computer parts and assembly price comparison.* Paper presented at the 2015 3rd International Conference on New Media (CONMEDIA).

Kesavaraj, G., & Sukumaran, S. (2013, 4-6 July 2013). *A study on classification techniques in data mining.* Paper presented at the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).

Kurte, K. R., Durbha, S. S., King, R. L., Younan, N. H., & Vatsavai, R. (2017). Semantics-Enabled Framework for Spatial Image Information Mining of Linked Earth Observation Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(1), 29-44. doi:10.1109/JSTARS.2016.2547992

Kuşak, L. (2014). Data Mining and Spatial Data Mining. *8*(1).

Lemieux, V., Rahmdel, P. S., Walker, R., Wong, B., & Flood, M. (2014). *Clustering techniques and their effect on portfolio formation and risk analysis.* Paper presented at the Proceedings of the International Workshop on Data Science for Macro-Modeling.

Li, S., Yang, Z., Xiu, J., & Liu, C. (2016, 26-28 Sept. 2016). *Research on an intelligent semantic-based information extraction framework.* Paper presented at the 2016 16th International Symposium on Communications and Information Technologies (ISCIT).

Löwe, R., Madsen, H., & McSharry, P. (2016). Objective Classification of Rainfall in Northern Europe for Online Operation of Urban Water Systems Based on Clustering Techniques. *Water, 8*(3), 87.

Mauricio, A. P., Payawal, J. M. M., Cueva, M. A. D., & Quevedo, V. C. (2016, 23-26 May 2016). *Predicting Customer Lifetime Value through Data Mining Technique in a Direct Selling Company.* Paper presented at the 2016 International Conference on Industrial Engineering, Management Science and Application (ICIMSA).

Mazhar, A., Ikram, M. T., Butt, N. A., & Butt, A. J. (2015, 2-4 Sept. 2015). *Do we really have to consider data mining techniques for meteorological data.* Paper presented at the 2015 Fourth International Conference on Aerospace Science and Engineering (ICASE).

Nelson, P. (2017). Natural Language Processing (NLP) Techniques for Extracting Information. Retrieved from <https://www.searchtechnologies.com/blog/natural-language-processing-techniques>

Pawlik, M., & Augsten, N. (2016). Tree edit distance: Robust and memory-efficient. *Information Systems, 56*, 157-173. doi:<https://doi.org/10.1016/j.is.2015.08.004>

Poch, M., Bel, N., Espeja, S., & Navio, F. (2014). *Ranking Job Offers for Candidates: learning hidden knowledge from Big Data.* Paper presented at the LREC.

Ramisa, A., Yan, F., Moreno-Noguer, F., & Mikolajczyk, K. (2018). BreakingNews: Article Annotation by Image and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(5), 1072-1085. doi:10.1109/TPAMI.2017.2721945

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., . . . Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664-681. doi:<https://doi.org/10.1016/j.neucom.2017.06.053>

Somvanshi, M., & Chavan, P. (2016, 12-13 Aug. 2016). *A review of machine learning techniques using decision tree and support vector machine.* Paper presented at the 2016 International Conference on Computing Communication Control and automation (ICCUBEA).

Song, Y., Huang, L., & Liang, S. (2009, 25-28 July 2009). *Research on rule definition and engine for general text processing.* Paper presented at the 2009 4th International Conference on Computer Science & Education.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2014). Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of Hydrology, 512*, 332-343. doi:<https://doi.org/10.1016/j.jhydrol.2014.03.008>

Wang, F., Wang, Q., Nie, F., Yu, W., & Wang, R. (2018). Efficient tree classifiers for large scale datasets. *Neurocomputing, 284*, 70-79. doi:<https://doi.org/10.1016/j.neucom.2017.12.061>

Wasilewska, A., & Menasalvas, E. (2018). A Granular Model for Data Mining.