Automated knowledge base Big Data repository for Enterprise domain using Ontology based Decision Support System

Yamuna Nagasandra Rajaiah

December 18, 2018

Abstract

In this era, enterprises are facing challenge in keeping track of large number of available Big data tool information. Domain experts with great technical expertise have been hired to drive business into right direction in terms of big data application scenarios for enterprise domain. Thus, this work begins with scraping data from heterogeneous data sources using ontology based approaches for data pre-processing using suitable tools and technologies and integrate information into repository. Scraping data from Internet using concept hierarchy takes another dimension which is not discussed in this paper. Tools for efficient scraping of HTML pages and text are very few for example Apache Nutch, Apache UIMA, Scrapy using python, Selenium using JAVA.The proposed model focuses on extracting information regarding big data tools like tool name, usage, version, system pre-requisites, compatibility and many more parameters. Starting with the web crawling technique employed to impersonate large amount of data from web by directing to a website, which would otherwise be difficult for a humans to extract. Firstly, data from various websites is taken as input, which is usually unstructured or semi-structured and the scraping program will convert data into machine readable structured format. The approach followed in this proposal would be (KDD) Knowledge Discovery in Databases modeling that includes processes like understanding of the application domain, understanding and creation of the dataset, data pre-processing, data transformation, evaluation and integration of the discovered knowledge to help in correct decision making. Result of this approach is building a repository of tools and version information for decision support system.

1 Introduction

The focus is to come up with a web-service platform to provide big data knowledge base repository in order to create value to the business decisions. Organizations have been involved in constant skill upgrade of the resources and investing time and effort for to taking business level decisions in order to achieve better results to select suitable tool for specific data use-cases in big data use cases. Selecting right tool based on the date usage is huge impact to organization. Implementation focuses on building tool information repository. The repository should provide the user with information on the big data tools and technologies that leverage on the decision making criteria for small and medium enterprises (SME). To retrieve relevant information based on the decisions made by the user on specified tool for this sort of implementation work, it is necessary that web scraper is able to aggregate as much big data tool information as possible from many third-party websites. The process is followed by applying knowledge discovery, a Data mining approach to extract data from websites and converting this data in machine readable format for further data processing steps like data cleaning, analyzing data and integrating with knowledge base for further visualization process. Domain experts would leverage the platform to take efficient decision with less duration than expected compared to legacy approach.

Big Data analytics has found itself as a rapidly growing technology and evolving as increasingly important in recent years for all industries as huge amount of data is being generated, analyzed and used at an extra-ordinary scale. [NHS16a] The big data tools that undertake processing of the data are typically programs that need to work on the complete datasets, and these programs make use of the computation on existing data by data mining models created, else recompute when new data are added to the datasets. [HRLHM15] The Correct selection of Big data tool is an huge challenge to different organizations and hence, they leverage on big data engineers with prominent

technical knowledge and experience to specify the best possible solution to yield better business decisions.Focus is to come up with a methodology to scrape data from multiple data sources to build a Big Data Knowledge based repository to make decisions based on KDD approach. The Main goal would be for better decision-making process based on Business interest of the organization, which would include increase profitable margins as well as retain existing customers and attract potential customers at different market places.

2 Research Question

The proposal is subjected towards exploiting discovery of knowledge. Process on Semantic web information and discover ontology approaches to build a Large Business Applications in the domain of Big data. Experts have knowledge of manually navigating through web sources to extract information, how could this be implemented automatically in terms of available Data Mining approach is the base for this research approach.

3 State of the Art

According to several researchers referenced in this paper, and considering the analysis and assumptions, Implementation of big data projects have ideal Hadoop technology as the priority. However, this could be false since different areas and organizations demands taxonomies which are different from one another and specific to targeted group. [SQ14] In terms of semantic web based approaches to gather data is driving more interest in the domain of knowledge discovery. Information extraction is the main idea to fetch syntactically and semantically relevant data. In case of extracting domain specific data, there is less focus given for semantic related data. Extracting statistically relevant data is more important in the projected approach.[ZL18]
Developing approaches may be to provide just a narrow scope of certain field or to leverage classification as Decision-support would be good choice. In terms of Ontology development, there can be need for classifying technologies according to individual phases of KDD (Knowledge Discovery in Databases) or the CRISP-DM (Cross Industry Standard Processes for Data Mining), these models results in specific approaches related to big data framework based on implementation use cases. [SQ14]

4 METHODOLOGY

Ontology plays a vital role in case of application scenarios that depends on formalization and its purpose and factors. One can observe many approaches and languages have been tried and tested to describe ontology and their behavior. Semantic based web search is necessary to collect as much of data as possible from various information available over Internet. To describe this more clearly in the domain of World Wide Web (WWW) to process data by machines to achieve those results. Initially documents need to be formatted in machine readable forms for example format data into (XML) Extensible Mark-up Language, (RDF) Resource Description Framework or (OWL) Ontology Web Language.[RP16] In recent years, to progress in development of ontologies where manual analysis and interpretation would be time consuming, various methodologies are considered which depend on target application domain, techniques and methods. Based on various research from academic learners as well as industry experts, work-flow of CRISP-DM and KDD methodologies are considered as a standard.

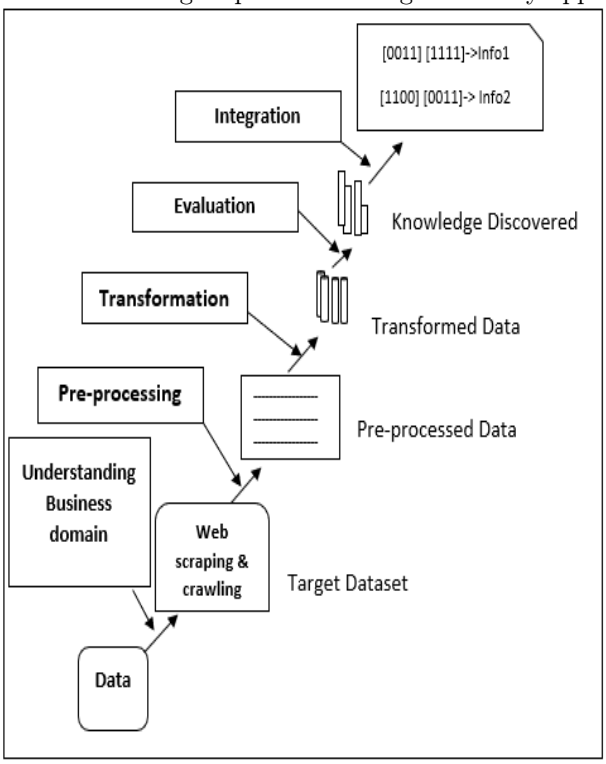
4.1 Overview

In this age, of Information industries, big data technology and projects are taking immense scope and always asks for expertise to solve critical problem which demands knowledge and experiences from Subject Matter Experts trained in particular field of knowledge. Solution to certain specific big data problem depends on the technology and need of the project use cases. Selecting perfect architecture for use cases are cumbersome, which requires ability to take decisions if organizations have nearly poor idea in this process.[Uk17]

4.2 Knowledge Discovery in Databases (KDD) comparison to CRoss Industry

Comparing the KDD stages with the CRISPDM, one can observe that the KDD methodology incorporates each phase in detail and tasks through official documentation and concrete examples by following steps that are as referred for Data Collection, Data Processing, Transformation and Knowledge discovery through specific tools and technologies. The KDD process including the understanding the business model phase can be identified with the understanding of the application domain, the relevant prior knowledge and the goals of the end-user. [MHAQ14] The Deployment phase can be consolidated by incorporating the discovered knowledge into the system. Concerning the remaining stages, one can say that, Data Understanding phase can be identified as the combination of Target dataset Selection and pre-processing. The Data Preparation phase can be acknowledged with Transformation and the Evaluation phase can be mapped with Data Interpretation. Thus, one can define the KDD approach is a process model used for extracting the hidden knowledge in the data according to requirement.

Figure 1: Processing steps of Knowledge discovery approach



4.3 Knowledge Discovery in database approach and its Implementation

There are 6 steps involved in this approach: 1. Data understanding in the business model 2. Understanding and creation of data set 3. Data pre-processing and cleaning 4. Data transformation 5. Evaluation of discovered Knowledge 6. Integration of discovered knowledge. Refer to Figure 1: Processing steps of Knowledge Data Discovery methodology followed in this implementation phase Each of the steps are discussed in detail in following sub-section:[NHS16b]

4.3.1 Step 1

Understanding of the Application domain includes the business analysis for comprehension of problem definition and project goal which includes data understanding. In this step, thorough facts and figures of the Big data application domain is required, relevant prior knowledge of organization data which utilizes the input and the accuracy of expert decision making by the big data tool is necessary.

4.3.2 Step 2

Understanding and creation of the dataset concerns with collecting and focusing on subset of data like Manufacturing company name, Tool version, System Requirement, Compatibility Information etc., from different big data vendors through web scraping tools and this data is used further for decision making process.

4.3.3 Step 3

Data pre-processing is a primary step for data discovery and focuses mainly to improve the quality of data by identifying irrelevant and inconsistent data. It includes data collection and integration, data techniques, where it is necessary to prepare the data through cleaning, data transformation and for knowledge discovery. Data pre-processing deals with transforming raw data extracted into an understandable format, as real-world data is incomplete and may contain errors.
Data Cleaning is a method to get accurate information by removing noise or outliers and strategies for handling inconsistency, so that complete data could be utilized as input for extracting big data knowledge from prepared data. Usually, the data to be analyzed may be incomplete i.e., can be lacking certain attributes of interest for application domain, may contain only aggregate data or noisy information which contains errors which differ from anticipated facts and can be inconsistent while categorizing details. Data Cleaning process would basically work on cleaning the data by identifying and eliminating outliers, resolving inconsistencies and smoothing noisy data. Although data cleaning may be done manually, but its is very time-consuming. Hence, it is feasible to can clean data interactively with different tools available or an integrated batch processing through scripting approach may be followed.

4.4 Step 4

Data Transformation: It is the process to transform data from one form to another to avoid invariant representation of data and it is necessary to get the required format for representation in the application domain of the project.

4.5 Step 5

Evaluation of the discovered knowledge: Before proceeding to the knowledge integration, it is necessary to check if the data is novel and interesting to interpret them by an expert domain and evaluate the impact of the discovered knowledge.

4.6 Step 6

Integration of the discovered knowledge: It consists of the integration and the deployment of the discovered knowledge from various big data web sources, which includes metadata information. Metadata here specify the forms of data available after the Data integration process, including fields such as company name, tool version, organization and download-able file. The possible values of fields are considered as finite. In terms of search queries of the form Microsoft should be able to retrieve solution related to the proprietary vendor and other information in the application.

4.7 Very Large business process model

Business process model refers to activities performed by humans to accomplish one or more business goals. Business models are generic and can occur in several sectors like marketing, health-care, financial management etc, hence, they generate large amount of data which develops new challenges on daily basis.Usually, the business process is combined with external data sources to ensure right execution of involved tasks through the adapted methodology. Functionality of the model is easily adaptable for large application since the steps focused in this approach is verified and tested. [HG17]
The Process involves four steps. These main phases of the model are implemented keeping several criteria as a reference. Initially the method is started by design phase. In this phase, raw data to be extracted is to be understood. Usually by representing the model based on Unified modeling approach. The next phase is to execution of the model. Several integration techniques need to be considered. This phase is followed by management which varies in terms of applications

used.Without the monitoring the process, it is not feasible to achieve maintainability of the cost effective models that has been built and then the methodology follows optimization that affects the performance and effectiveness of the processed model.
The process have been explained in the following sub-section:

4.8 Design

The process is based on understanding of application domain and execution of graphical representation using use-case diagrams to show the interactions between users and system involved.

4.9 Execution

Post design of the process, model would be integrated with external domain sources

4.10 Management and supervision

The process is deployed in execution environment, which is constantly monitored and managed.

4.11 Visual Analysis and Optimization

The Collected data can be analyzed through visual representation, which makes it easier in identifying areas which show poor performance and that need least attention.

5 Feasible Experimental Results

Based on the comparison of available approaches like CRISP-DM and KDD which operates on different design phases, a well know KDD approach was selected to develop a framework resulting in classifying domain specific information. [NJ97] The demonstrated process model is capable of building a system which is more accurate and cost-effective than domain knowledge experts. As discussed in the paper, the decision support system is utilizing different Knowledge Discovery approaches to process the data with keen interest of providing efficient results in the business decision making helping organizations who have little idea about big data tools. In order to overcome the risk of inaccurate classification of the domain knowledge one can leverage the platform to drive customer data into proper direction.

6 Conclusion and Future work

In recent years, the business environment has come across different challenges due to high volume of data stored and requirement to enhance efficiencies in decision making. In this paper, the discussion about feasible ontology-based decision support system which assists organizations to provide solutions for business-critical decisions relevant to big data domain. Initially in this work, a brief overview of the current scenario and need for specific model was discussed. After that based on the literature research conducted, present state of the art for classifying big data technologies and need for new approach was justified.
Ontology based approaches are examined thoroughly and applied in the implementation workflow, in order to enable the end-users to get a thorough understanding of certain hidden knowledge which may result in profitable growth of companies there is a requirement to develop a web-based application platform using existing various open-source tools and technologies.
In the future, it is expected to find methodologies that is capable of extracting information from social media websites and twitter data analysis and also can consider domain specific information extraction techniques for example in case of health-care and e-commerce platforms. Other than platforms one can think of extracting data from Cloud which required another level of data processing model and need to consider issues with respect to processing data in term of network models supported by Cloud Infrastructure. Extracting information based on keyword search requires expertise in Natural language processing (NLP) techniques. Retrieving information based on the linguistic and ranking process is considered to be efficient.

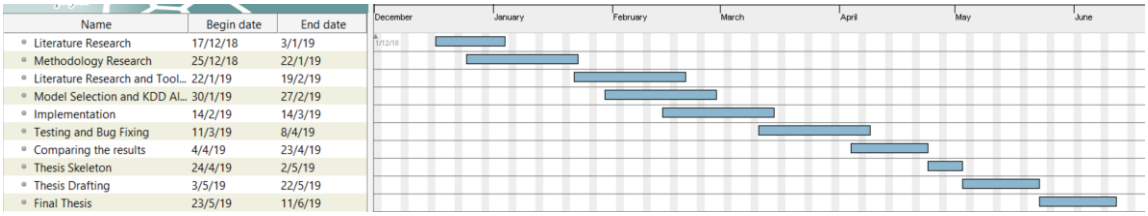


Figure 2: Project Plan for 24 weeks

References

[HG17] Asma Hassani and Sonia Ghannouchi. A framework for business process data management based on big data approach. *Procedia Computer Science*, 121:740–747, 01 2017.

[HRLHM15] M Misbachul Huda, Dian Rahma Latifa Hayun, and Zhin Martun. Data modeling for big data. *Jurnal ULTIMA InfoSys*, 6:1–11, 12 2015.

[MHAQ14] Mortadha M Hamad and Banaz Anwer Qader. Data pre-processing for knowledge discovery. *Tikrit Jourual of Pure Science*, 19:143–148, 01 2014.

[NHS16a] Michael. Chui James. Manyika Tamim. Saleh Bill. Wiseman Nicolaus. Henke, Jacques. Bughin and Guru. Sethupathy. 2016.

[NHS16b] Michael. Chui James. Manyika Tamim. Saleh Bill. Wiseman Nicolaus. Henke, Jacques. Bughin and Guru. Sethupathy. 2016.

[NJ97] Sambasivan Narayanan and Sundaresan Jayaraman. *A knowledge-based decision support system for apparel enterprise evaluation*, pages 67–110. 01 1997.

[RP16] Petar Ristoski and Heiko Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, 36, 01 2016.

[SQ14] Umair Shafique and Haseeb Qaiser. A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12:2351–8014, 11 2014.

[Uk17] Ijeacs Uk. Survey on big data analytics. *International Journal of Engineering and Applied Computer Science (IJEACS)*, 02:181–185, 07 2017.

[ZL18] Xianyi Zeng and Jie Lu. Decision support systems with uncertainties in big data environments. *Knowledge-Based Systems*, 143:327, 03 2018.