Framework Based Ontology for Heterogenous Big Data Correlation in Cloud Infrastructure

Tengku Adil Tengku Izhar  
Faculty of Information Management  
Universiti Teknologi MARA  
Puncak Perdana, UiTM Selangor, Malaysia  
tengkuadil@yahoo.co.uk

Bernady O. Apduhan  
Faculty of Information Science  
Kyushu Sangyo University,  
Fukuoka, Japan  
boa93dai@gmail.com

Abstract—People have access to more data in single day than most people that have access to data in the previous decade. This data is created in many forms and it highlights the development of Big Data. Big Data in organizations have transformed the way organizations across industries implement new approach to handle huge amount of data. Organizations rely to this data to achieve specific business priorities. The challenge is how to capture this data from different data sources. The aim of this paper is to propose a framework using an ontology to capture relevant information from different data sources. In order to achieve this aim, we review past literature related to the model based ontology that incorporate both internal data and external data sources. The results from the this paper will develop an incremental approach to extend the ontology with relevant external knowledge bases by clustering and annotating the external data to concepts and properties within the ontology.

Keywords—Big Data; Cloud; Enterprise; Ontology; Decision-making

I. INTRODUCTION

Big data provides significant opportunities for organizations to impact a wide range of practices and processes in the organization. However, there is still little debate these days on the role and importance of big data for efficient decision-making. While most people take such technologies for granted, our understanding about external data such as social media, mobile data and sensor data are still very limited. Most existing studies on this topic only focus on knowledge extraction from each data source independently, and the outcomes are then combined and co-relation is investigated [1, 2]. However, in this proposed research we argue that these external data sources need to be semantically linked and integrated before any meaningful data analytics or knowledge extraction can be performed. Furthermore, we proposed a framework to find a correlation between internal and external data for efficient decision-making. Although there are studies on statistics correlations and machine learning based correlations, but as to our knowledge, there is a lack of study on using a semantic underpinning such as an ontology to perform correlation of heterogenous data in a cloud infrastructure.

Study on cloud shows cloud storage architecture based on ontology provide comprehensive capabilities to store and retrieve files in the cloud based on its contents [4]. According to [5], there are usually several cloud computing services/resources with different features being deployed by the same or different organizations on cloud environments. This leads the need to support interoperability between different cloud services/resources to capture data from large databases. Although there are studies that focus on cloud and ontology development for managing organizational data and information [8], there is still lack of evidence which focus on the development of ontology to provide semantic connectivity between internal organizational data with external sources for efficient decision-making. The aims of this research are:

1. To propose a framework that underpins a seamless integration of organizational data and heterogeneous external data pertinent to the organizational focus area.
2. To develop a mechanism that allows seamless consolidation of knowledge from external sources
will enrich the capability of the organization to make accurate decision-making.

The availability of heterogeneous external sources are growing very significantly in the last few years, especially due to the availability of wireless and mobile technologies, crowd-sourcing facilities, Internet of Things, sensor networks, and other social media and web data. The focus of our research is to generate huge amount of data that can be extracted to generate values to the organization and to establish situational awareness of the community or market trends.

II. LITERATURE REVIEW

Organizations have at their disposal a large volume of data with a wide variety of types. Technology-driven organizations want to capture process and analyze this data at a fast velocity for efficient decision-making to support their operations and their business processes [3-6]. As much as data volume and variety increases and as faster analytic results are needed, more demanding is for data architecture. This data architecture should enable collecting, storing, and analyzing Big Data [7]. Many companies such as Amazon, Google and Microsoft accelerate their paces in developing cloud-computing systems and enhancing their services to cater for a wide variety of users [8].

The bulk of existing academic literature on social media has been published in just the last few years and has focused on the social processes of social media and its effects in areas such as marketing, politics, health communication, and education [9]. Social media platform such as Twitter has stormed onto the social media scene not only as an individual communication device but also as an information dissemination platform [10]. People on social media express opinions on different topics. There are few studies performing external data competitive analysis on the leading companies in an industry in a systematic way [11]. External data can help decision makers to ensure efficient solutions to the problems raised [12]. However, the trustworthiness of this data is often questionable due to the huge amount of data created. For example, social media has now become an important medium of communication and interaction tools for social networks [13]. Social media is also important for business platform that can influence the corporative environment [14-16]. Social networks involve agents in creating and processing information for knowledge network [17]. At the same time, the role that causality can play in social network analysis is unclear [18]. Therefore, it is important to examine the flow of information share on social media and to retrieve relevant information from large amount of it. While most people take such technologies for granted, our understanding about external data such as social media, mobile data and sensor data are very limited [1]. Most existing studies on this topic remain descriptive, focusing on what people do with it [1, 2]. Despite the abundant research on IT adoption in general, our understanding about the effectiveness of analyzing external data is still at the early stage. Moreover, the uniqueness of this data from the other IT applications may require for information entrepreneur to further their theoretical extension.

A. Cloud for information enterprise

Progress of research efforts in a novel technology is contingent on having a rigorous organization of its knowledge domain and a comprehensive understanding of all the relevant components of cloud technology and their relationships [19-22]. Study by Youseff et al. [19] showed that cloud computing is one contemporary technology in which the research community has recently embarked. The technology would enable the organization to design more efficient portals and gateways for analyzing their data and facilitate the adoption of this novel computing approach in scientific environments. For example, cloud storage architecture based on ontology can store and retrieve files in the cloud based on its content [20].

Cloud computing requires scalable and cooperative sharing the resources in various organizations by dynamic configuring a virtual organization according to user’s requirements. Study by Yoo et al. [21] showed that ontology-based representation of cloud computing environment would be able to conceptualize common attributes among cloud resources.

Retrieving data from a well organized database is requested to be familiar with its schema, structure, and architecture. An effective approach to retrieve desired information or to extract useful data is an important issue in the emerging cloud [22]. For example, in the mobile cloud environment, the information retrieval system based on the P2P is very important. In case of an information search that is not consistent with the meta profile in the data distribution technique, the reliability of the searched data cannot be assured [23]. According to Surachai & Banditwattanawong [24], there is usually several cloud computing platforms with different features being deployed in the same or different organizations in a multi-cloud environment. This leads to the need to interoperate between different cloud platforms to capture data from large databases.

Cloud concept is still changing as a large pool of easily usable and accessible virtualized resources. These resources can be dynamically for an optimum resource utilization [25]. For example, McFedries [26] described data center as the basic unit of cloud. It offers huge amounts of computing power and storage by using spare resources, which relate to the concept of massive data scalability [27]. Most of ongoing works are aiming at developing the techniques and constructing cloud platforms, such as Amazon, Google AppEngine and Microsoft Azure [28]. Massive data in data center of cloud platform can provide benefits to the cloud provider and consumer in retrieving information among business, medical information, and cooperative information
retrieval platform [29, 30]. Table 1 shows limited study on cloud that focus on retrieving information from external data. In contrast to our study, we develop ontology on cloud for information retrieval. Based on our observation only Surachai & Banditwattananwong [24] and Yoo et al. [21] incorporate ontology and cloud.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Ontology</th>
<th>Information retrieval</th>
<th>Information</th>
<th>External data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chang et al. [22]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Golovichinsky et al. [29]</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Surachai &amp; Banditwattananwong [24]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>McFedries [26]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Yoo et al. [21]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Chow et al. [30]</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

B. Ontology for information enterprise

Most of the recent studies on ontology focus on system ontology and enterprise ontology. Some studies include both system ontology and enterprise ontology while some studies focus on organizational ontology. In Rao et al. [31], the authors discussed ontology as representing organizational knowledge which provides the means by which to understand the relationships between the organizational goals, sub-goals, business processes, tasks, sub-tasks, resources and decision makers. In this study, the development of ontology is based on knowledge mapping within the enterprise modelling and the author identifies the flow of knowledge within the organization.

On the other hand, Sharma & Osei-Bryson [32] presented an organization-ontology-based framework that not only incorporates the applicable tools and techniques, it also provides the ability to present the output of activities in a form that allows for at least their semi-automated integration with activities of this phase and succeeding phases. However, the authors look at the data mining methodology that is associated with business understanding (BU). This study was followed up by Rao et al. [33], who proposed an approach by building an ontology and a set of corresponding competency questions for the information technology (IT) infrastructure domain. The authors also emphasized that a formal set of ontologies must have a set of formal axioms that provide the basis of the ontology’s deductive capability. Mansingh et al. [34] proposed an ontology-driven methodology for extracting different knowledge items and representing them as knowledge maps.

Kang et al. [35] deliberated on enterprise architectures (EA) as an approach to address the problems between humans or between systems or between humans and systems. In order to solve this problem, the authors developed an ontology based on enterprise architecture. Kang et al. [36] also developed a business enterprise ontology and identified the lack of semantics which causes communication problems between humans or between systems or between humans and systems. The authors used a fact-based ontology as a conceptual modelling method to cope with the dynamically changing business environment.

These days, the activities of enterprises are continuously globalized and the business environment is changing rapidly and becoming more complex [37]. In response to the changing business environment, it is important to develop new business models and business processes. Kang et al. [35] developed an ontology based on enterprise modelling to maintain competitiveness by accommodating changes in the business environment quickly and flexibly. However, Kang, Lee & Choi [35] and Kang, Lee & Kim [36] focused on system development based on ontology. Even though these studies are based on an ontology model, the authors do not emphasize any relationship between the internal data and external data in their model.

Jimeno-Yepes et al. [38] studied ontology refinement to improve information retrieval. The authors studied the ontology and terminological resources which have appeared in information retrieval (IR) either to provide query expansion terms, to perform semantic indexing of documents or to assist in the better organization of retrieved documents. However, this ontology is usually not optimal for IR tasks. Han & Park [39] studied an enterprise ontology where knowledge is treated as a critical driving force for attaining the organization’s performance goals which is important for decision-making. They proposed a knowledge framework model and an enterprise ontology for the process-centered organizational structure. Huang & Dao [40] focused on the usage of ontology to integrate the knowledge between different domains to improve business processes. The authors believe that organizations are becoming more knowledge intensive, and the integration of various types of knowledge becomes a challenge. Based on this study, an ontology-based workflow is developed to accumulate knowledge in on-going processes and can provide help in complex workflow systems and optimization. Table 2 shows that there is no study focus on cloud that integrates ontology to retrieve information. In this study, we propose ontology on cloud to store and retrieve information from external sources such as social media and mobile data.

Table 2. Summary of the previous studies on ontology.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Cloud</th>
<th>Information retrieval</th>
<th>Information</th>
<th>External data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rao et al. [31],</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Huang &amp; Dao [40]</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Jimeno-Yepes et al.</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
III. FRAMEWORK

Previously, a study has been done on organizational data and social media data using an ontology to capture relevant information and to resolve the issue in identifying and evaluating relevant data for better decision-making that covers the characteristics of good quality relevant data, Fig. 1. In the current research, we proposed an organization Global Ontology that can capture the concepts of complexity and use it as a basis for data integration processes suitable for efficient correlative decision-making. We will develop an incremental approach to extend the ontology with relevant external knowledge bases by clustering and annotating the external data to concepts and properties within the ontology.

The originality of this research lies in the creation of the big data integration ‘road-map’ in the form of an organization Global Ontology that contains references to concepts and properties from external sources relevant to the organization. This road-map will enable accurate integration of the datasets for decision-making as it will serve as a meta-data for the connectivity of the integration process. This approach will enable the organization to have an overall view of data connectivity within and outside the organizations, and to enable the data scientists to harvest interconnected information for analytics purposes.

This organization Global Ontology will reside on the cloud and initially contain the local organizational ontology. The on-premise storage services will house the organizational data and the different off-premise storage services will house the external data which we collectively consider the big data. Our first task is to develop the local organizational ontology to show the flow of data in the organization. After that, we will focus to develop the relationship between local ontology and external data. This local ontology will be extended by clustering the external data to define the degree of relevance of the particular data. The organization Global Ontology will extract knowledge through correlation between local ontology and external data.

![Fig. 1. Proposed semantic-based Big Data correlation on cloud.](image-url)
address which data from the datasets are more important. The outcome of this research can establish an analytics of Big Data structure for the organizations to ensure that analytics processes are supported by the specific organizational priorities. The contribution of an ontology is to improve the creation of model ultimately takes place through different domain.

In Fig. 1, the Organization Global Ontology is built on a cloud computing environment. The on-premise storage services will house the organizational data and the different off-premise storage services will house the external data, which we consider the Big Data. This Big Data is basically dynamic and its relation/connectivity to the organizational data needs to be incrementally and timely updated based on-demand from the user. Organization Global Ontology will contain the local organizational ontology and it will be extended by clustering and annotating the relevant external data in a timely manner. The first focus of this research will be on the development of the local organizational ontology to show the flow of data in the organization. After that, we will focus on the relationship between local ontology and external data. This local ontology will be extended by clustering the external data to define the degree of relevance of the particular data. In this research, Organization Global Ontology will extract knowledge through the correlation between local ontology and external data. In doing correlation, there are few issues that we want to resolve. For example, finding the most efficient clustering methodology to extract external data. Likewise, it is important to find the best way to link the local organizational ontology to the matched data, which was derived after clustering the external data.

IV. Expected Results and Impacts

Impact on Society – Many workplaces in general creates huge amount of data. This study will improve the productivity in terms of knowledge management implication when dealing with large volume of data. Such workplaces will not only practice values of managing data, information and knowledge but also to promote trustworthiness of data and information in multi-disciplinary background that advocates these values to the society.

Impact on the Economy – Having an ability to manage data is crucial ingredient for k-economy that helps to transform a part of human knowledge to machines by using particular knowledge technology. This technology will improve performance in dealing with data, work performance and efficient decision-making delivery for more competitive and productive economy. Therefore, having the ability to analyze data in a timely fashion can ensure the ability to improve productivity in relation to decision-making. This research is multi-disciplinary that could find correlation between data in different domains such as disaster management and people movements, biology and agriculture sciences, health care and population, etc.

Impact on the Nation - The results can lead to some new evaluation methods in big data era from different perspectives. The outcome will offer an enormous opportunity to advance the science of data analytics so that future researchers will have a new understanding on what is needed to improve their data analysis process. The conduct of this research will ensure the projects agility in responding to unfolding events, and substantially enhance its ability to engage in and impact on societies and nation.

V. Conclusion

Social media such as Facebook and Twitter have become an indispensable part of our lives. In this paper, we have described the main features of ontology when developed the relationship between internal data and external data. We extend the application of the ontology to evaluate data from different data sources. This paper enables researchers to classify and evaluate existing research, to design scientific research, to identify the gaps and weaknesses prior to future research directions.

REFERENCES


