Jurnal Teknologi

Full Paper

THE ARCHITECTURE OF INFORMATION EXTRACTION FOR ONTOLOGY POPULATION IN CONTRACTOR SELECTION

Rosmayati Mohemad[a*], Abdul Razak Hamdan[b], Zulaiha Ali Othamn[b], Noor Maizura Mohamad Noor[a]

[a]Software Technology Research Group (SofTech), School of Informatic and Applied Mathematics, Universiti Malaysia Terengganu, 21030, Kuala Terengganu, Terengganu, Malaysia
[b]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bandar Baru Bangi, 43600, Selangor, Malaysia

Article history
Received
20 October 2015
Received in revised form
24 October 2015
Accepted
1 August 2015

*Corresponding author
rosmayati@umt.edu.my

Abstract

The enormous amount of unstructured data presents the biggest challenge to decision makers in eliciting meaningful information to support business decision-making. This study explores the potential use of ontologies in extracting and populating the information from various combinations of unstructured and semi-structured data formats such as tabular, form-based and natural language-based text. The main objective of this study is to propose an architecture of information extraction for ontology population. Contractor selection is chosen as the domain of interest. Thus, this research focuses on the extraction of contractor profiles from tender documents in order to enrich ontological contractor profile by populating the relevant extracted information. The findings are significantly good in precision and recall, in which the performance measures have reached an accuracy of 100% precision and recall for extracting information in both tabular and form-based formats. However, the precision score of relevant information extracted in natural language text is average with a percentage of 42.86% due to the limitation of the linguistic approach for processing Malay texts.

Keywords: Decision-making, information extraction, ontology population, contractor selection

1.0 INTRODUCTION

In a recent International Data Corporation (IDC) Digital Universe study, the volume of unstructured digital data was reported to have increased exponentially where the amount had grown to 1,227 exabytes in 2010 and it is forecasted to grow by 45.2% to 7,910 exabytes in 2015 [1]. Sources of such unstructured data are found in very different formats including, but not limited to documents, emails, video, image, social media, etc. The proliferation of high dimensionality of unstructured data presents one of the key challenges to decision makers in eliciting meaningful information to support business decision-making processes. Unstructured data contains hidden knowledge that benefits the decision makers in accessing information for decision selection [2].

Spanning across multiple domains of knowledge, the emergence of researches on ontology has facilitated the decision-making process, such as in the domain of environment [3, 4], supply chain [5], financial [6, 7] and transport [8]. Meanwhile, in the field of information extraction, ontology has been exploited to extract relevant information and populate the output in the ontology from tabular data [9, 10], form-based data [11, 12] and textual data [13, 14]. However, to the best of our knowledge, none of these researches had exploited the role of ontology in assisting information extraction as one of the integrative component in supporting decision-making since data management methods developed for structured data are not directly applicable. Therefore, this study explores the potential use of ontologies in extracting and populating information from various combinations of unstructured and semi-structured data formats such

as tabular, form-based and natural language-based text.

Ontology is a formal representation of knowledge enriched with language semantics, that enables the creation of a set of concepts, properties and individuals, which understood by both humans and machines. It does not only serve to encapsulate and represent knowledge about some domain of interest, but it is also capable of processing the content of information and reasoning about it. Many past studies have adopted ontology for knowledge management, including knowledge modeling, knowledge sharing, knowledge reuse, knowledge discovery and knowledge acquisition in a wide range of domains such as construction [15, 16], manufacturing [17], medical [18, 19] and education [20-22].

The main objective of this study is to propose an architecture of information extraction for ontology population. To demonstrate the feasibility of the proposed architecture, the contractor selection process is chosen as the domain of research. The selection of the best contractor through tendering processes plays a pivotal role in the success of a construction project [23]. Details and comprehensive information on contractor profiles are essential for a transparent, fair and reliable assessment of contractor qualification. In order to achieve this, contractors are required to provide various information in tender documents, normally in free-format, to support their qualification [24, 25]. However, these unstructured and ill-defined formats of tender documents make the contractor selection process difficult. Thus, this research focuses on the extraction of contractor profiles from documents in order to enrich the ontological contractor profile by populating the relevant extracted information, for the purpose of the best contractor selection.

The rest of the paper is organized as follows. First, the Related Works section presents the state-of-the-art information extraction tools, paying attention to the ontology-based information extraction. Meanwhile, the subsequent section outlines the overall framework of the research methodology. The proposed architecture of information extraction for ontology population is discussed next. The evaluation results are explained in the following Results and Discussion section. Finally, the last section concludes with a summary of this paper and future research directions.

2.0 RELATED WORKS

Information extraction applies different types of approaches (e.g. ontology, knowledge discovery, pattern matching) to derive unstructured and semi-structured data stored in machine-readable documents into a structured format, so that it can be effectively manipulated by the computer [26]. The main task of information extraction is to analyze relevant data containing information appropriate to the task at hand. Many previous studies have explored the use of ontology in assisting the information extraction process. Ontology plays four main tasks in the context of information extraction, which are: (1) ontology as a guide, (2) ontology as a repository, (3) ontology as a representation schema, and (4) ontology as a basis for reasoning [27].

Ontology-based information extraction is characterized as: (1) a system to process natural language text in unstructured (free-form text) or semi-structured formats (i.e. tabular, form-based or web-based), (2) with guidance from pre-existing ontology to identify salient information, and (3) to populate the extracted information as a set of instantiated and related concepts and attributes in the ontology [28]. Based on these characterizations, two main tasks of ontology are emphasized in this research, which will be discussed further in this section. Firstly, the task of ontology is as a guide to detect salient information. The second task of ontology is as a repository where the extracted information is stored through ontology population. Output from the ontology population is useful for further processing in the context of decision-making, information retrieval, information, integration, knowledge construction, knowledge acquisition and others.

Table 1 summarizes the comparison of ontology-based information extraction systems from previous researches in terms of its input text formats (unstructured or semi-structured) and outputs produced from the extraction process. AJAX data extraction is a web-based extractor tool that employs agricultural ontology and extraction algorithms to guide information extraction and annotation from dynamic AJAX content [29]. Here, the proposed framework only works for semi-structured data and ignores combined natural language processing text.

Table 1 Summary of ontology-based information extraction

Extractor Tool	Input Format	Output
AJAX Data Extraction [29]	Semi-structured	Individuals
WebOMSIE [12]	Semi-structured	Values
FLOPPIES [30]	Semi-structured	Individuals
T2FOBOMIE [31]	Semi-structured	Values
ORP [11]	Unstructured	Individuals
CCG-IE [14]	Unstructured	Concepts Individuals

Another extractor tool that deals with semi-structured data is WebOMSIE, an ontology-based information extraction from heterogeneous Web sources [12]. WebOMSIE uses ontology to infer the approach that is adequate for the extraction according to the Web sources involved. The tool however requires users to manually pre-define the characteristics of information extraction criteria such as document format, URL, page number and others. Meanwhile, Ali *et al*. [31] proposed a system called the type-2 fuzzy ontology-based opinion mining and

information extraction (T2FOBOMIE). The system adopted ontology and type-2 fuzzy mechanism to extract user's desires by reformulating user query for hotel reservation. Another information extraction application was proposed by Nederstigt *et al*. [30] to extract product information from tabular data sources on the e-commerce Web. Table is a common form of presenting data and often contains factual information that could be useful for better product comparison or recommendation by exploiting the semantic of the products' attributes and the corresponding values.

In processing unstructured data, Faria *et al.* [14] developed an extractor tool, called the CCG-IE system, which is an ontological structure in the domain of molecular biology that guides the information extraction process from scientific articles. Several analyses were done for small corpus of text in abstracts and full articles. Whilst in human resource management, Younsi *et al*. [11] adopted ontology to assist in the extraction of information from millions of resumes in Microsoft Word documents for the purpose of employment. They proposed the Ontology-based Resume Parser (ORP). ORP represents the semantic data in a resume, such as personal information, business and academic, experience, skills, publications, certifications and others. The tool is capable of detecting significant information from pre-asserted data through semantic based inference rules.

However, all these researches have only examined a single document format, either free-form text (unstructured) or web-based text (semi-structured), and have not considered the semantic of document format in combinations of unstructured and semi-structured data. In addition, the significant results of information extraction output were not evaluated further for the purpose of ontology population by demonstrating the usefulness in other applications. In this research, we not only exploit ontology to assist in the information extraction process for both unstructured and semi-structured data, but we also demonstrate the use of enriching ontology through ontology population in supporting decision-making based on a contractor selection case study.

3.0 RESEARCH METHODOLOGY & APPROACH

This section describes the overall framework of the research methodology carried out in this study. As depicted in Figure 1, the framework consists of three main phases, including the development of the ontology model, the development of the extraction engine and evaluation.

PHASES	METHODS	OUTPUT
Ontology Model Development	• Literature study • Domain experts interviews • Document analysis	Application Ontology
Extraction Engine Development	• Experimental approach	Information Extraction for Ontology Population Architecture
Evaluation	• Accuracy measurement of precision, recall & f-measure	Evaluation of extraction performance

Figure 1 The framework of research methodology

3.1 Ontology Model Development Phase

Developing the ontology model is a knowledge engineering process, which refers to the activities involved in building the ontology process. To date, various approaches of knowledge engineering have been applied to build the ontology model [32]. Methontology as proposed by [33], is taken as the knowledge engineering methodology in this study, which includes five steps.

The first step is knowledge specification where it is intended to determine the purpose and scope of ontology by considering its benefits supporting the current application and potential extensibility in the future. Different domain ontologies may have different purposes and may apply different methods. For instance, the main purposes of the ontology model developed in this study are to facilitate information extraction and to support the decision-making process. Literatures of related domain are reviewed and analyzed, as well as reusable candidate ontologies are identified and evaluated in this step. Besides, a series of unstructured interviews with domain experts are conducted in order to gain insights and a better understanding of the domain. Document analysis is another method used in this step to study the format, structure and content of the documents involved. The main outcome is the informal perception of a domain, represented in a set of ontology requirement specification.

The second step is knowledge conceptualization. This is the most important step as it determines the rest of the ontology development. The main task of this step is to organize and structure the knowledge acquired from the previous step into a conceptual model using any independent knowledge representation paradigms such as UML or semantic networks. The knowledge conceptualization

depends on the purpose and scope of the ontology building. At the end of this step, all the necessary ontology components such as concepts, attributes, relations, formal axioms, rules and individuals of the domain, together with their glossary of terms are constructed. For example in this study, three conceptual models are structured accordingly: (1) the document structure represents three different data formats such as tabular, form-based and natural-based text, (2) the key indicators used to capture the relevant information for contractor selection domain are structured as the conceptual model, and (3) the contractor profiles. However, detailed discussion of these conceptual models is beyond the scope of this paper.

Meanwhile, the third step is knowledge integration. Concepts and semantic relationships in other existing domain ontologies are taken into consideration to be reused and expanded by integrating the knowledge when constructing the model. The fourth step is knowledge implementation, which is a process of interpreting knowledge into the targeted representation language so that it is readable by the computer. The selection of ontology implementation language depends on the expressiveness and reasoning capabilities of the ontology languages. Here, the conceptual model (concepts, relations, attributes, formal axioms, rules and individuals) defined in the second and third steps are coded into the Web Ontology Language (OWL) format. OWL is a standard language based on the World Wide Web Consortium (W3C). It is published by using the syntactic Resource Description Framework (RDF), and the schematic language for RDF is Resource Description Framework Schema (RDFS). Many ontology editors have been developed to support knowledge implementation such as TopBraid Composer and Protégé. The last step is knowledge maintenance. The content of constructed ontology needs to be evaluated by the domain experts. Therefore, any additional knowledge and changes are updated and corrected in the ontology during the maintenance activity.

3.2 Extraction Engine Development Phase

A sequence of operations is executed by the adopting the experimental approach to develop the extraction engine. The details of the experimental design are presented in Figure 2. This phase serves to extract relevant information from unstructured (natural language-based text) and semi-structured (tabular and form-based) documents, and then to populate the extracted information into ontology. The design can be divided into four main steps.

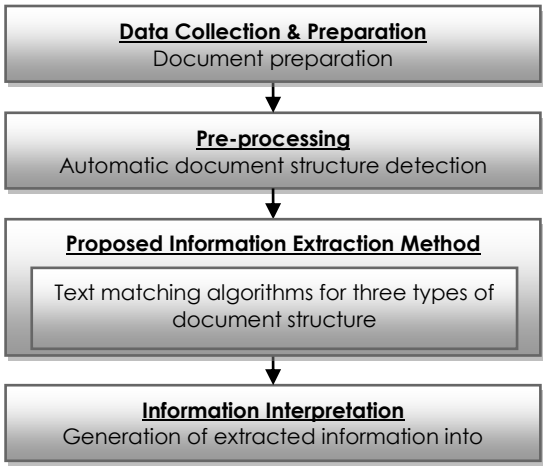
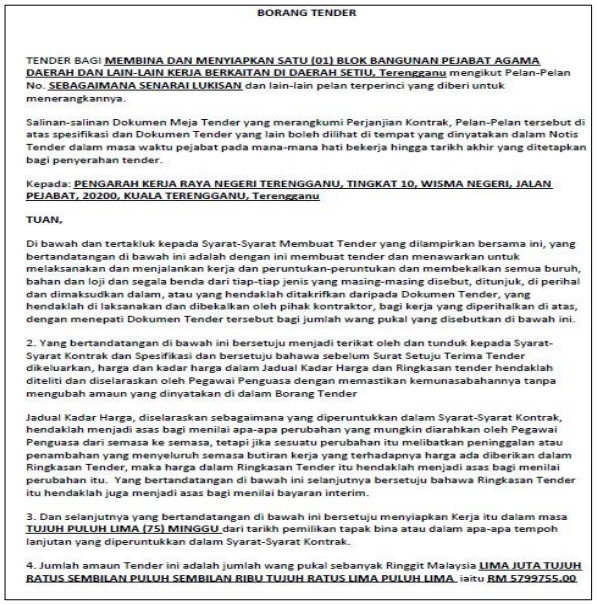


Figure 2 The experimental design

The first step consists of data collection and preparation. In this step, a set of tender data is compiled. The data contains contractor profiles and tender information. The reason for this sample selection is due to its information-rich, alphanumeric document, which is arranged in various structures either in unstructured or semi-structured forms and delivered in the Malay language. The corpus used in this study is collected from real world tender documents for building constructions from the Malaysian Public Work Department. The tender documents were originally in hardcopy and were transformed into the Portable Document Format (PDF). Each document consists of various contractor profile information that is visually represented in tabular, form-based and natural language-based text formats. Figure 3 and Figure 4 depict the sample of unstructured and semi-structured tender documents respectively.



BORANG TENDER

TENDER BAGI **MEMBINA DAN MENYIAPKAN SATU (01) BLOK BANGUNAN PEJABAT AGAMA DAERAH DAN LAIN-LAIN KERJA BERKAITAN DI DAERAH SETIU, Terengganu** mengikut Pelan-Pelan No. **SEBAGAIMANA SENARAI LUKISAN** dan lain-lain pelan terperinci yang diberi untuk menerangkannya.

Salinan-salinan Dokumen Meja Tender yang merangkumi Perjanjian Kontrak, Pelan-Pelan tersebut di atas spesifikasi dan Dokumen Tender yang lain boleh dilihat di tempat yang dinyatakan dalam Notis Tender dalam masa waktu pejabat pada mana-mana hati bekerja hingga tarikh akhir yang ditetapkan bagi penyerahan tender.

Kepada: **PENGARAH KERJA RAYA NEGERI TERENGGANU, TINGKAT 10, WISMA NEGERI, JALAN PEJABAT, 20200, KUALA TERENGGANU, Terengganu**

TUAN,

Di bawah dan tertakluk kepada Syarat-Syarat Membuat Tender yang dilampirkan bersama ini, yang bertandatangan di bawah ini adalah dengan ini membuat tender dan menawarkan untuk melaksanakan dan menjalankan kerja dan peruntukan-peruntukan dan membekalkan semua buruh, bahan dan loji dan segala benda dari tiap-tiap jenis yang masing-masing disebut, ditunjuk, di perihal dan dimaksudkan dalam, atau yang hendaklah ditakrifkan daripada Dokumen Tender, yang hendaklah di laksanakan dan dibekalkan oleh pihak kontraktor, bagi kerja yang diperihalkan di atas, dengan menepati Dokumen Tender tersebut bagi jumlah wang pukal yang disebutkan di bawah ini.

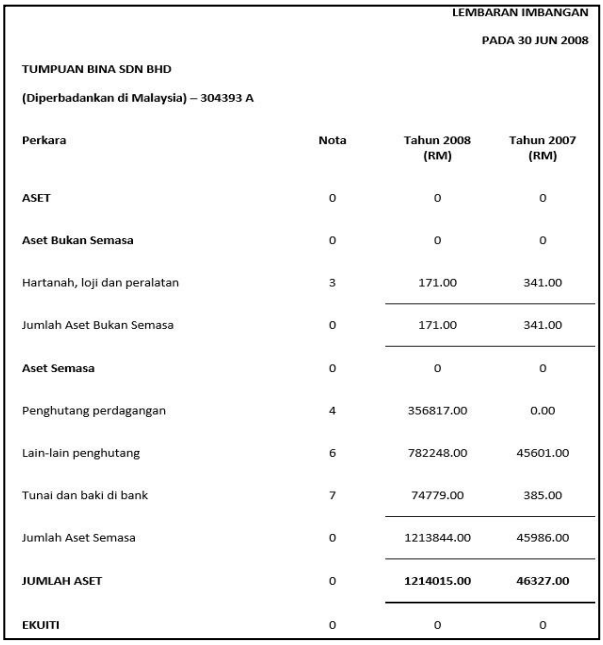
2. Yang bertandatangan di bawah ini bersetuju menjadi terikat oleh dan tunduk kepada Syarat-Syarat Kontrak dan Spesifikasi dan bersetuju bahawa sebelum Surat Setuju Terima Tender dikeluarkan, harga dan kadar harga dalam Jadual Kadar Harga dan Ringkasan tender hendaklah diteliti dan diselaraskan oleh Pegawai Penguasa dengan memastikan kemunasabahannya tanpa mengubah amaun yang dinyatakan di dalam Borang Tender

Jadual Kadar Harga, diselaraskan sebagaimana yang diperuntukkan dalam Syarat-Syarat Kontrak, hendaklah menjadi asas bagi menilai apa-apa perubahan yang mungkin diarahkan oleh Pegawai Penguasa dari semasa ke semasa, tetapi jika sesuatu perubahan itu melibatkan peninggalan atau penambahan yang menyeluruh semasa butiran kerja yang terhadapnya harga ada diberikan dalam Ringkasan Tender, maka harga dalam Ringkasan Tender itu hendaklah menjadi asas bagi menilai perubahan itu. Yang bertandatangan di bawah ini selanjutnya bersetuju bahawa Ringkasan Tender itu hendaklah juga menjadi asas bagi menilai bayaran interim.

3. Dan selanjutnya yang bertandatangan di bawah ini bersetuju menyiapkan Kerja itu dalam masa **TUJUH PULUH LIMA (75) MINGGU** dari tarikh pemilikan tapak bina atau dalam apa-apa tempoh lanjutan yang diperuntukkan dalam Syarat-Syarat Kontrak.

4. Jumlah amaun Tender ini adalah jumlah wang pukal sebanyak Ringgit Malaysia **LIMA JUTA TUJUH RATUS SEMBILAN PULUH SEMBILAN RIBU TUJUH RATUS LIMA PULUH LIMA** iaitu **RM 5799755.00**

Figure 3 Samples of unstructured tender document



LEMBARAN IMBANGAN

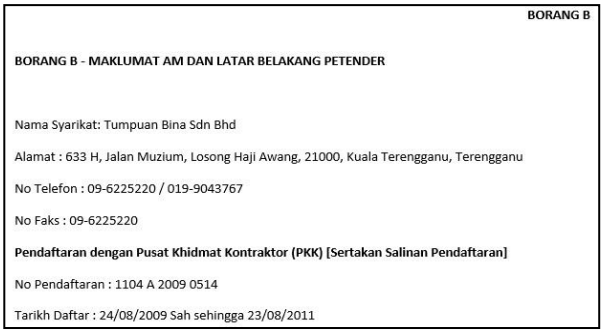
PADA 30 JUN 2008

TUMPUAN BINA SDN BHD

(Diperbadankan di Malaysia) – 304393 A

Perkara	Nota	Tahun 2008 (RM)	Tahun 2007 (RM)
ASET	0	0	0
Aset Bukan Semasa	0	0	0
Hartanah, loji dan peralatan	3	171.00	341.00
Jumlah Aset Bukan Semasa	0	171.00	341.00
Aset Semasa	0	0	0
Penghutang perdagangan	4	356817.00	0.00
Lain-lain penghutang	6	782248.00	45601.00
Tunai dan baki di bank	7	74779.00	385.00
Jumlah Aset Semasa	0	1213844.00	45986.00
JUMLAH ASET	0	**1214015.00**	**46327.00**
EKUITI	0	0	0

(a)



BORANG B

BORANG B - MAKLUMAT AM DAN LATAR BELAKANG PETENDER

Nama Syarikat: Tumpuan Bina Sdn Bhd

Alamat : 633 H, Jalan Muzium, Losong Haji Awang, 21000, Kuala Terengganu, Terengganu

No Telefon : 09-6225220 / 019-9043767

No Faks : 09-6225220

Pendaftaran dengan Pusat Khidmat Kontraktor (PKK) [Sertakan Salinan Pendaftaran]

No Pendaftaran : 1104 A 2009 0514

Tarikh Daftar : 24/08/2009 Sah sehingga 23/08/2011

(b)

Figure 4 Samples of semi-structured tender document; a) tabular b) form-based

The second step is pre-processing where the document structure is detected automatically. In order to automate this process, the machine is capable of recognizing the relevant text and associating it with the appropriate document structure. Figure 5 illustrates the series of activities in pre-processing of the PDF documents. Here, every text in the document is parsed to obtain its structure through tokenization, distance estimation, clustering and structure detection rules. These activities are inspired and modified from the study conducted by Oro and Ruffolo [34]. The details of this approach was discussed by Rosmayati *et al*. [35].

Meanwhile, the third step is the process of automatically extracting relevant tender information. Text matching algorithms for three different types of document structure are developed to trace the relevant information in the document based on the key indicators ontology and document structure ontology constructed in the ontology development phase. The last step is information extraction where the list of extracted information is populated as individuals into the contractor profile ontology. Definition of semantic knowledge in the ontology is used to associate an individual with the concept. The process can be seen as a step-by-step ontology learning.

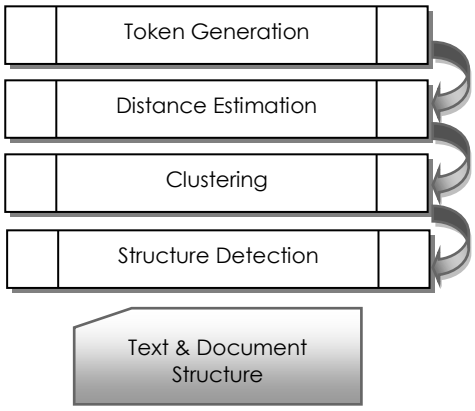


Figure 5 Series of pre-processing activities for pdf document

3.3 Evaluation Phase

The most important phase in research methodology is the evaluation of the proposed architecture. The objective of the evaluation in this study is to measure the accuracy of the information extraction process. Thus, the evaluation is conducted to determine the accuracy of the extractor in extracting relevant contractor profiles in tender documents. This section explains the approach in detail.

In order to measure the accuracy, standard information extraction methods of precision, recall and f-measure are used. Precision is defined as the proportion of retrieved information that is relevant, while recall is the proportion of relevant information that is retrieved. f-measure combines the precision and recall scores. Below are the formulas applied for these three measurements.

$$Precision = \frac{R}{R+1} \quad (1)$$

$$Recall = \frac{R}{R+N} \quad (2)$$

$$f-measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

R represents the number of information identified as relevant information, I is the number of irrelevant information identified and N refers to the number of relevant information failed to be identified.

3.0 THE DESIGN ARCHITECTURE

This section discusses the architecture of information extraction for ontology population. The architecture is characterized by five major components namely, extraction engine, storage engine, access engine, inference engine and knowledge base. These five components are the mechanism that assists information extraction from unstructured and semi-structured documents to support decision-making in contractor selection. Figure 6 represents the architecture proposed in this study. The source data is a set of information-rich tender documents in PDF.

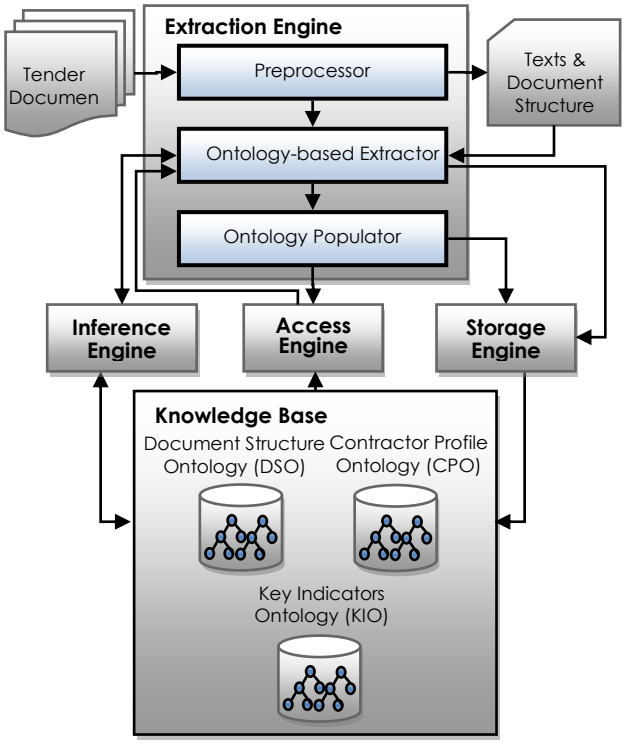


Figure 6 The architecture of information extraction for ontology population

4.1 Extraction Engine

The extraction engine is where the actual ontology-based information extraction and ontology population processes take place. The main function of this component is the extraction of contractor profile information from tender documents. It is supported by three sub-components: preprocessor, ontology-based extractor and ontology populator.

The preprocessor is necessary to reduce the high dimensionality problem of processing textual and numerical data in unstructured and semi-structured documents. As depicted in Figure 6, the input of an extraction engine first goes through a preprocessor, which converts the textual and numerical data to a format that can be handled by the ontology-based extractor. It first tokenizes sequences of consecutive characters into larger units, called token. Then, the resulting tokens are clustered based on their distance measures. This process is necessary to group some tokens to determine the document structure. Output resulting from the preprocessor is a list of text with the description of its document structure.

The second sub-component is the ontology-based extractor, which carries out the extraction of possible relevant contractor profiles based on the output produced by the preprocessor. Examples of contractor profiles are contractor background, bidding price, bidding project completion time, financial data, technical staff, list of construction plant and equipment, as well as past and current project performances. In order to perform this operation, a corresponding list of text is firstly annotated using matching algorithms based on Key Indicators Ontology (KIO). A rule-based extraction approach is adopted in this sub-component, in which several rules are created to search for the possible relevant contractor profile information. Below is the example of the rule to infer the bidding price (BP) from the possible extracted contractor information.

$$BP = \forall Paragraph \cap \exists.\, hasKeyword\ KeyBidPrice \cap \exists.\, hasKeyword\ KeyMonetary \cap \exists.\, isConsisted\ FormTender \quad (4)$$

There are three outcomes resulting from the ontology-based extractor: (1) possible extraction of contractor profile in the unstructured form (natural language-based text), (2) possible extraction of contractor profile in the semi-structured form (tabular), and (3) possible extraction of contractor profile in the semi-structured (form-based). These outcomes are populated as individuals in the Document Structure Ontology (DSO) using the ontology populator.

The ontology populator involves the extraction and classification of individuals into the relevant concepts in the pre-defined ontology. Two main tasks are carried out by the ontology populator. Firstly, it populates the outputs identified by the ontology-based extractor in the DSO. Secondly, it serves to interpret the possible contractor profile extracted previously and populates the relevant contractor profile as individuals in the Contractor Profile Ontology (CPO). The process of populating in this study is essential for supporting decision-making in selecting the most qualified contractor as the populated ontology contains structured data that can be queried by the decision-making application.
The ontology-based extractor and the ontology populator interact with ontologies in the knowledge base through the inference, access and storage engines. The ontologies that are used by both components are generated separately by an ontology editor.

4.2 Inference Engine

The inference engine is executed from the Java API. It is used for the purpose of extraction rules execution in translating semantic information for a statement or fact defined explicitly in the knowledge base.

4.3 Access Engine

The access of semantic knowledge contained in the knowledge base is conducted through the access engine using the processing query. The access method used in this study is based on the SPARQL protocol and the RDF query language. The ontology-based extractor and the ontology populator achieve the necessary knowledge through this query interface.

4.4 Storage Engine

Storage engine serves to store explicit knowledge models in the form of tripplestore in the knowledge base. This engine allows the definition of the conceptual model.

4.5 Knowledge Base

Knowledge base consists of application ontologies for a specific domain of interest. Since the selected domain is contractor selection, thus in this study, three different ontologies are developed: Key Indicators Ontology (KIO), Document Structure Ontology (DSO), and Contractor Profile Ontology (CPO).

KIO includes the concepts and relations of key indicators concerning contractor profiles. Figure 7 shows the snapshot of the KIO, which lists the key indicators for annotating possible contractor profiles such as tender title, bidding price and bidding time to complete the construction in an unstructured text. Meanwhile, the DSO represents the semantic knowledge of tender document structures. As mentioned previously, text in tender documents is either in unstructured or semi-structured forms. The ontology consists of three main concepts, which represents the document structures: tabular, form-based and non-tabular. Figure 8 explains the snapshot of the DSO. In order to support th decision-making process for selecting the most eligible contractor, concepts and semantic relations are represented in the Contractor Profile Ontology (CPO). Therefore, the population of individuals from the extracted information in CPO serves to support the decision-making application. Figure 9 depicts the snapshot of the CPO.

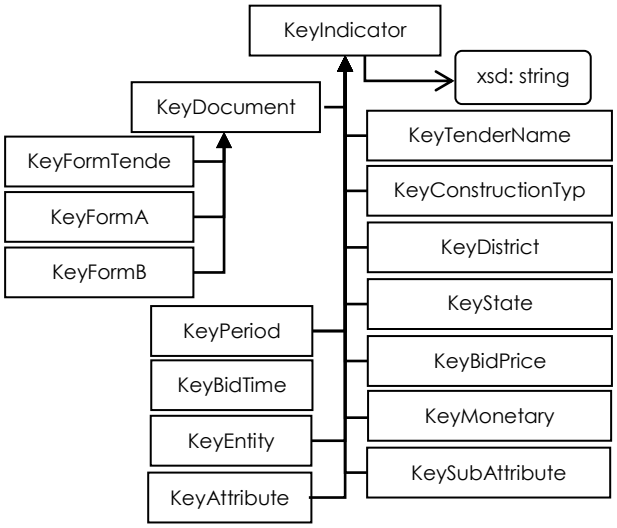


Figure 7 Snapshot of key indicators ontology (KIO)

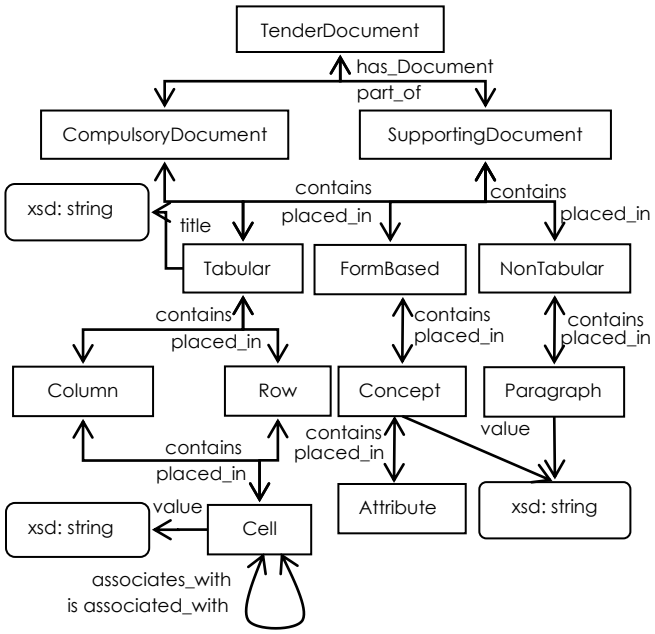


Figure 8 Snapshot of document structure ontology (DSO)

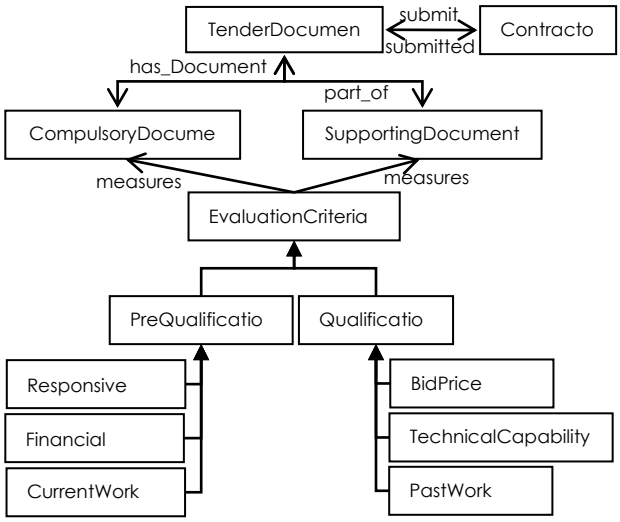


Figure 9 Snapshot of contractor profile ontology (CPO)

3.0 RESULTS AND DISCUSSION

The performances of the proposed architecture are evaluated based on the performance of the ontology-based information extractor system for contractor selection. This section further discusses the evaluation results of this study. The purpose of this experiment is to evaluate the ability of the ontology-based information extractor based on precision, recall and f-measure scores. In this experiment, six copies of tender documents of similar building construction projects based on the Malaysia Construction Tender are used as the experimental data. The average pages per document is approximately fifteen page. Figure 10 shows the example of the natural text output for extracting bidding price by populating matched keyword into ontology. Meanwhile, Figure 11 and Figure 12 represents the examples of output for extracted information in both tabular and form-based format.

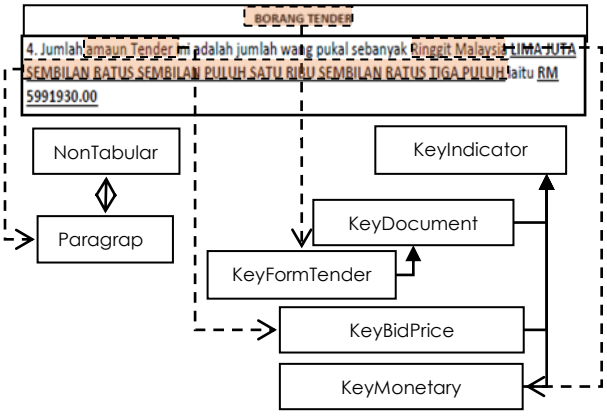


Figure 10 Example of extracted information in natural text

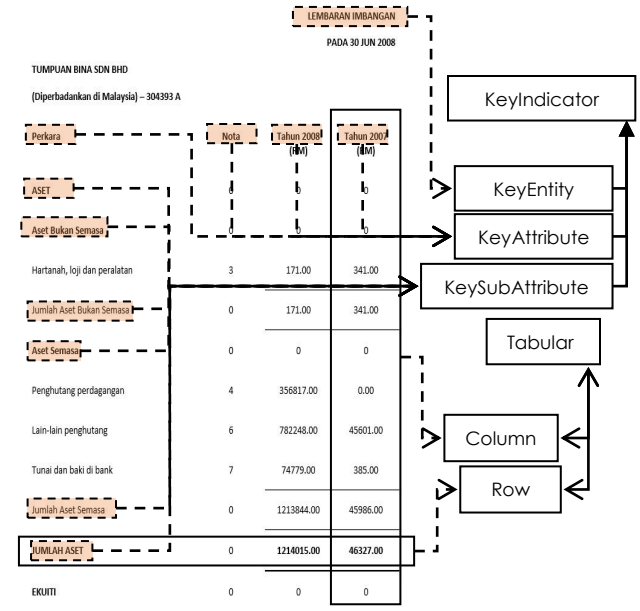


Figure 11 Example of extracted information in tabular

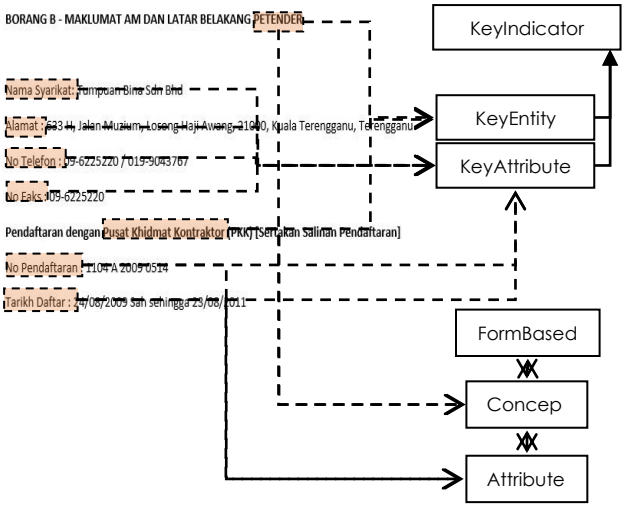


Figure 12 Example of extracted information in form-based

The ontology-based information extractor system successfully extracted the required information in tabular, form-based and natural text. The number of information extracted by the system is recorded in Table 2, Table 3 and Table 4. The figures are arranged according to the document structure respectively. The comparison is done between the actual data present in the tender documents and the data retrieved by the ontology-based information extractor system. Based on Table 2, the number of relevant and irrelevant information in the tabular format is identified based on table (T), column (C), row (R) and cell (L). Furthermore, as can be seen in Table 3, the comparison of information extracted in the form-based format is done based on the number of relevant and irrelevant concepts (C) and

attributes (A) retrieved. The concepts denote form title and the attributes represent the form fields. Meanwhile, Table 4 shows the number of relevant and irrelevant information retrieved in the natural text.

Table 2 Information extracted in tabular compared to actual data

Tender Document	Tabular											
	# Actual Relevant Items Detected				# System Relevant Items Detected				# System Irrelevant Items Detected			
	T	C	R	L	T	C	R	L	T	C	R	L
D1	29	129	189	903	29	129	189	903	0	0	0	0
D2	26	143	144	781	26	143	144	781	0	0	0	0
D3	35	185	200	1041	35	185	200	1041	0	0	0	0
D4	26	124	145	706	26	124	145	706	0	0	0	0
D5	17	81	105	483	17	81	105	483	0	0	0	0
D6	25	106	157	699	25	106	157	699	0	0	0	0

Table 3 Information extracted in form-based compared to actual data

Tender Document	Form-based					
	# Actual Relevant Items Detected		# System Relevant Items Detected		# System Irrelevant Items Detected	
	C	A	C	A	C	A
D1	4	18	4	18	0	0
D2	3	10	3	10	0	0
D3	4	18	4	18	0	0
D4	4	18	4	18	0	0
D5	3	10	3	10	0	0
D6	4	18	4	18	0	0

Table 4 Information Extracted in Natural Text Compared to Actual Data

Tender Document	Natural Text		
	# Actual Relevant Items Detected	# System Relevant Items Detected	# System Irrelevant Items Detected
D1	3	3	4
D2	3	3	4
D3	3	3	4
D4	3	3	4
D5	3	3	4
D6	3	3	4

Meanwhile, the average precision, recall and f-measure scores of the overall performance of the information extractor are shown in Table 5. The evaluation of precision, recall and f-measure shows significantly good accuracy in detecting relevant information in tabular and form-based formats, whilst the performance score for identifying information in the natural text-based is average. The precision rates for tabular, form-based and natural text are 100%, 100% and 42.86% respectively. The recall has achieved 100% for tabular, form-based and natural text. Meanwhile, the tested accuracy of the f-measure are 100 % for both tabular and form-based, and 60.00% for natural text.

Table 5 Average precision, recall and f-measure scores

	Precision	Recall	f-measure
Tabular	100.00%	100.00%	100.00%
Form-based	100.00%	100.00%	100.00%
Natural Text	42.86%	100.00%	60.00%

The average score of precision in natural text is due to the limitation of rules constructed. Since this study involves information in Malay text, it limits the use of the linguistic approach in the information extractor system. This is because the current researches in the part-of-speech tagging, stemming and lemmatization for processing the Malay language are still immature. However, the good accuracy results for information extraction in tabular and form-based formats show that the ontology is significantly capable of recognizing important information in the semi-structured format. It suggests that the structure information defined by key indicators in ontology can be used to recognize relevant text of the content of a tender document. Nevertheless, this finding is significant for non-complex tabular structures (basic table with single layer header and body) with complete definition of key indicators.

4.0 CONCLUSION

The architecture of information extraction for ontology population in the contractor selection domain is supported using a keyword-based approach, as modelled in the ontology. The use of keywords is essential to ensure only necessary contractor profile information is extracted and populated. The precision, recall and f-measure are proposed to strengthen the results of information extraction. Based on the findings, the extracted information has shown significant results in terms of precision, recall and f-measure for the extracted information in the semi-structured format. However, the keywords and rule-based implementations in this study allow any matched information especially in the natural language-based text to be annotated because it does not consider the linguistic view. There is a possibility of unnecessary recognition of information and ambiguous interpretation of

knowledge. In future, the meanings of information will be morphologically analyzed using a linguistic approach before information extraction is carried out. Furthermore, table recognition which currently works based on a simple table assumption, can be enhanced into a complex table.

Acknowledgement

This research is fully supported by FRGS grant. The authors fully acknowledged Ministry of Higher Education (MOHE) and Universiti Malaysia Terengganu for the approved fund which makes this important research viable and effective.

References

[1] Gantz, J. F., and Reinsel, D. 2011. The 2011 Digital Universe Study: Extracting Value from Chaos. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[2] Power, D. J., Burstein, F., and Sharda, R. 2011. Reflections on the Past and Future of Decision Support Systems: Perspective of Eleven Pioneers. Decision Support: An Examination of the DSS Discipline. Schuff, D., Paradice, D., Burstein, F., Power, D. J. and R. Sharda eds. Springer New York. 25-48.

[3] Wanner, L., Rospocher, M., Vrochidis, S., Johansson, L., Bouayad-Agha, N., Casamayor, G., Karppinen, A., Kompatsiaris, I., Mille, S., Moumtzidou, A., and Serafini, L. 2015. Ontology-Centered Environmental Information Delivery for Personalized Decision Support. *Expert Systems with Applications*. 42(12): 5032-5046.

[4] Hou, S., Li, H., and Rezgui, Y. 2015. Ontology-based Approach for Structural Design Considering Low Embodied Energy and Carbon. *Energy and Buildings*. 102(2015): 75-90.

[5] Wang, X., Wong, T. N., and Fan, Z.-P. 2013. Ontology-based Supply Chain Decision Support for Steel Manufacturers in China. *Expert Systems with Applications*. 40(18): 7519-7533.

[6] Chowdhuri, R., Yoon, V. Y., Redmond, R. T., and Etudo, U. O. 2014. Ontology Based Integration of XBRL Filings for Financial Decision Making. *Decision Support Systems*. 68 (2014): 64-76.

[7] Shue, L.-Y., Chen, C.-W., and Shiue, W. 2009. The Development of an Ontology-Based Expert System for Corporate Financial Rating. *Expert Systems with Applications*. 36(2): 2130-2142.

[8] Niaraki, A. S., and Kim, K. 2009. Ontology Based Personalized Route Planning System Using a Multi-Criteria Decision Making Approach. *Expert Systems with Applications*. 36(2): 2250-2259.

[9] Bahulkar, A., and Reddy, S. 2013. Ontology Driven Information Extraction from Tables Using Connectivity Analysis. Lecture Notes in Computer Science. R. Meersman, H. Panetto, T. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. De Leenheer, and D. Dou eds. Springer Berlin Heidelberg. 642-658.

[10] Guo, C., Ma, S., and Yuan, D. 2014. A Web Table Extraction Method Based on Structure and Ontology. Advanced Data Mining and Applications, Lecture Notes in Computer Science. Luo, X. Yu, J. and Li, Z. eds. Springer International Publishing. 695-704.

[11] Younsi, Z., Quafafou, M., Ouzegane, R. and Tari, A. 2013. WebOMSIE: An Ontology-Based Multi Source Web Information Extraction. *New Trends in Databases and Information Systems, Advances in Intelligent Systems and Computing*. Pechenizkiy, M. and Wojciechowski, M. eds. Springer Berlin Heidelberg. 199-208.

[12] Çelik, D. and Elçi, A. 2013. An Ontology-based Information Extraction Approach for Résumés. Pervasive Computing and the Networked World, Lecture Notes in Computer Science. Zu, Q., Hu, B. and Elçi, A. eds. Springer Berlin Heidelberg. 165-179.

[13] Moreno, A. Isern, D. and López Fuentes, A. C. 2013. Ontology-based Information Extraction of Regulatory Networks from Scientific Articles with Case Studies for Escherichia Coli. *Expert Systems with Applications*. 40(8): 3266-3281.

[14] Faria, C., Serra, I. and Girardi, R. 2014. A Domain-Independent Process for Automatic Ontology Population from Text. *Science of Computer Programming*. 95(Part 1): 26-43.

[15] Lu, Y., Li, Q., Zhou, Z. and Deng, Y. 2015. Ontology-based Knowledge Modeling for Automated Construction Safety Checking. *Safety Science*. 79(2015): 11-18.

[16] El-Diraby, T. E. and Osman, H. 2011. A Domain Ontology for Construction Concepts in Urban Infrastructure Products. *Automation in Construction*. 20(8): 1120-1132.

[17] Chungoora, N., Young, R. I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N. A., Cutting-Decelle, A.-F., Harding, J. A., and Case, K. 2013. A Model-driven Ontology Approach for Manufacturing System Interoperability and Knowledge Sharing. *Computers in Industry*. 64(4): 392-401.

[18] Bright, T. J. 2009. Development and Evaluation of an Ontology for Guiding Appropriate Antibiotic Prescribing. PhD. School of Arts and Sciences, Columbia University, Columbia.

[19] David, S., Arantza, A. and Clare, M. 2011. An Ontology of Diabetes Self Management. *Proceedings of the First International Workshop on Managing Interoperability and Complexity in Health Systems*. Glasgow, Scotland. 28 October 2011. 83-86.

[20] Jia, H., Wang, M., Ran, W., Yang, S. J. H., Liao, J. and Chiu, D. K. W. 2011. Design of a Performance-Oriented Workplace e-Learning System using Ontology. *Expert Systems with Applications*. 38(4): 3372-3382.

[21] Kuo-Kuang, C., Chien, I. L. and Rong-Shi, T. 2011. Ontology Technology to Assist Learners' Navigation in the Concept Map Learning System. *Expert System Application*. 38(9): 11293-11299.

[22] Fernandez-Breis, J. T., Castellanos-Nieves, D., Hernandez-Franco, J., Soler-Segovia, C., Robles-Redondo, M. d. C., Gonzalez-Martinez, R., and Prendes-Espinosa, M. P. 2012. A Semantic Platform for the Management of the Educative Curriculum. *Expert Systems with Applications*. 39(5): 6011-6019.

[23] Rosmayati, M., Abdul Razak, H., Zulaiha, A.O. and Noor Maizura, M. N. 2011. Modelling Ontology for Supporting Construction Tender Evaluation Process. *International Conference on Semantic Technology and Information Retrieval (STAIR 11)*. Putrajaya, Malaysia. 27-29 June 2011. 282-288.

[24] Ciribini, A. L. C., Bolpagni, M., and Oliveri, E. 2015. An Innovative Approach to e-Public Tendering Based on Model Checking. *Procedia Economics and Finance*. 21 (2015): 32-39.

[25] Schaaffkamp, C. 2014. How Can Customer Focus be Strengthened in Competitive Tendering? *Research in Transportation Economics*. 48(2014): 305-314.

[26] Small, S. and Medsker, L. 2014. Review of Information Extraction Technologies and Applications. *Neural Computing and Applications*. 25(3-4): 533-548.

[27] Sen, S., Tao, J. and Deokar, A. 2015. On the Role of Ontologies in Information Extraction. *Reshaping Society through Analytics, Collaboration, and Decision Support, Annals of Information Systems*. Iyer, L.S. and Power, D. J. eds. Springer International Publishing. 115-133.

[28] Daya, C. W. and Dejing, D. 2010. Ontology-based Information Extraction: An Introduction and a Survey of

Current Approaches. *Journal Information Science*. 36(3): 306-323.

[29] Li, C.-x., Su, Y.-r., Wang, R.-j., Wei, Y.-y., and Huang, H. 2012. Structured AJAX Data Extraction Based on Agricultural Ontology. *Journal of Integrative Agriculture*. 11(5): 784-791.

[30] Nederstigt, L. J., Aanen, S. S., Vandic, D., and Frasincar, F. 2014. FLOPPIES: A Framework for Large-Scale Ontology Population of Product Information from Tabular Data in e-Commerce Stores. *Decision Support Systems*. 59(2014): 296-311.

[31] Ali, F., Kim, E., and Kim, Y.-G. 2015. Type-2 Fuzzy Ontology-based Opinion Mining and Information Extraction: A Proposal to Automate the Hotel Reservation System. *Applied Intelligence*. 42(3): 481-500.

[32] Gomez-Perez, A., Corcho-Garcia, O., and Fernandez-Lopez, M. 2004. *Ontological Engineering*. Springer-Verlag New York.

[33] Fernandez, M., Gomez-Perez, A., and Juristo, N. 1997. METHONTOLOGY: From Ontological Art towards Ontological Engineering. *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering.* Menlo Park, California. 24-26 March 1997. 33-40.

[34] Oro, E. and Ruffolo, M. 2009. PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents. 10th International Conference on Document Analysis and Recognition. Barcelona, Spain. 26-29 July 2009. 906-910.

[35] Rosmayati, M., Abdul Razak, H., Zulaiha, A.O., and Noor Maizura, M. N. 2011. Automatic Recognition of Document Structure from PDF Files. *Software Engineering and Computer Systems, Communications in Computer and Information Science*. Zain, J. M., Wan Mohd, W. M. b., and El-Qawasmeh, E. eds. Springer Berlin Heidelberg. 274-282.