Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

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

OBJECT RECOGNIZER FOR ORGANIZING AND STRUCTURING UNSTRUCTURED DOCUMENTS USING INTERROGATIVE KNOWLEDGE

[1]ISKANDAR ISHAK, [2]FATIMAH SIDI, [3]MARZANAH A. JABAR

[1]Senior Lecturer, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
[2]Assoc. Prof., Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
[3]Assoc. Prof., Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

E-mail: [1]iskandar_i@upm.edu.my, [2]fatimah@upm.edu.my, [3]marzanah@upm.edu.my

Knowledge in unstructured documents is lacking the structured characteristics; therefore raising the problem of extracting answers when queried by simple queries. This paper proposes the Interrogative Knowledge Object- (IKO-) Recognizer that is able to model the extracted interrogative lexical constructs from unstructured documents into ontological constructs. An experiment is carried out to test performance of the IKO-Recognizer using quantitative retrieval performance metrics of recall and precision. The result shows that the extracted interrogative constructs is able to increase human understanding on knowledge available in unstructured document.

Keywords: *Interrogative Knowledge, Unstructured Documents, Object Recognizer, Information Retrieval*

1. INTRODUCTION

Structured document always have a well-defined hierarchical structured arranged or organized in rows and columns. However, this is not the case for the content of unstructured documents where it is not in organized manner. Lacking of structured characteristics in such documents has made it difficult for human to extract knowledge to answer queries, even with an automatic system such as knowledge discovery in databases [1]. Embley *et al.* [2, 3, 4, 5] establish and develop an approach of extracting information from unstructured documents and reformulate the information as relations in a database. Later, [6] extend the use of information extraction using ontology approach that leads to semantic understanding based on a foundation of Medow's definitions [7] for data, information, knowledge and meaning.

In the current explosion of Big Data, data comes from multiple resources in the forms of databases, social media and sensors. Large part of the data such blow post, social media status and postings data are in unstructured format or textual format which is in the form of sentences. This has lead to investigation on transforming information in unstructured documents into a structured form by imposing some structuring characteristic over the content of the unstructured document. This will enable database to query using the standard Data Manipulation Language (DML), hence facilitating human understanding on the overall content of the unstructured document.

This study proposes a solution to organize knowledge in interrogative nature from the unstructured documents by transforming the extracted knowledge into interrogative structured form. This is based on the integration of interrogative-based [8] and conceptual modelling [2, 3, 4, 5, 6] with further enhancement on deep-level understanding of complete sentences. It refers to the understanding of a group of words in a complete sentence which begin with a capital letter and end with a full stop, question mark, or exclamation mark. The ultimate goal is to capture knowledge in the knowledge-base system and

database management system via interrogative knowledge organization and structuring.

Philosophical thoughts on ontology are the metaphysical study on the basic categories and relationships of being and existence based on conceptualization [9]. From the computer science perspective, ontology is a theory of data model that represents the objects in a particular domain with relationships between the object. It is a specification of concepts defining set of representational terms [10] that associates names of entities in the universe of discourse of classes, attributes and the relationships that may exist between those concepts. Besides, ontology allows a community to agree upon the meaning of terms and relations so that they may reliably reuse and share the knowledge [11, 12].

Researchers over the past years have adopted a number of approaches and methods on the transformation of information to data through hierarchy [13, 14], ontology [2, 3, 4, 5, 6], database schema [15], frames [16], and maps [17, 18, 19, 20, 21, 22, 23] and lexical database such as WordNet [24]. However, to achieve maximum benefits, the Conceptual-Modelling approach based on ontology has an added advantage. The data extraction method based on conceptual modelling of OSM (Object-oriented System Model) established by Embley *et al.* [2, 3, 4, 5] and [6] have reported recall ratios in the range of 90% and precision ratios near 98% in extracting data on unstructured documents that are data rich. A constant/keyword recognizer uses matching rules generated by a parser to extract and structure data. Specifically, their approach consists of the following five steps.

i. Develop an ontology model instance over an area of interest;
ii. Parse ontology to generate a database scheme and to generate rule for matching constants and keywords;
iii. Invoke a record extractor that divides an unstructured web document into individual record-size chunks, cleans by removing mark-up language tags and presents as individual unstructured record document for further processing;
iv. Invoke recognizers that use the matching rules generated by the parser to extract from the cleaned individual unstructured documents, the objects expected to populate the model instance; and
v. Populate the generated database scheme by using heuristics to determine which constants populate which records in the database scheme.

The Open Knowledge Base Connectivity (OKBC) has high expressiveness for ontology representation. It is a protocol to access knowledge in Knowledge Representation Systems (KRSs) like ontology repositories or object-relational databases. There are a number of ontology tools available such as Protégé, Ontolingua, Chimaera, OntoEdit, and Oiled. Among all tools, Protégé is found to have knowledge-modelling structures that are compatible with OKBC [25], hence very much suitable for creating concepts and relationship when not much reasoning support is available.

This paper advocates the use of ontology as knowledge representation towards the formation of knowledge organization and structuring in supporting creation of concepts or classes, properties, attributes, and relationships that may exist in unstructured documents. The ontology will be implemented using the Protégé-Frames editor due to the availability of classes, slots, and instances to organize and structure knowledge in hierarchical form as represented in the domain concepts.

The remainder of this paper is organized as follows. Section II will present the proposed Interrogative Knowledge Object Recognizer to perform interrogative knowledge organization and structuring. Section III will discuss the experimental results and finally Section IV will conclude this paper with some indication for future work.

2. PROPOSED IKO-RECOGNIZER

To cater the process of organizing and structuring interrogative knowledge, this paper proposes the Interrogative Knowledge Object- (*IKO*-) Recognizer, which is an object recognizer that is used to model the ontological constructs from the extracted interrogative lexical constructs. At the end, the lexical constructs will be presented in the form of objects with ontological relations of the extracted knowledge.

The *IKO*-Recognizer consists of two major processes, object recognizer and mapping process. First, the object recognizer uses object interrogative analysis rules by utilizing Object-Oriented Programming (OOP) in order to conceptually organize the program around its data

(objects/concepts). In this process, a number of object interrogative analysis rules and precondition language is pre-defined but users may manually define additional rules. Second, the following mapping process uses an ontology engineering approach, whereby objects that have been created by the object recognizer are accessible as plug-ins in the ontology system. Fig. 1 highlighted the main processes in the *IKO*-Recognizer, which are the Object Interrogative Analysis Rules and the Precondition Language.

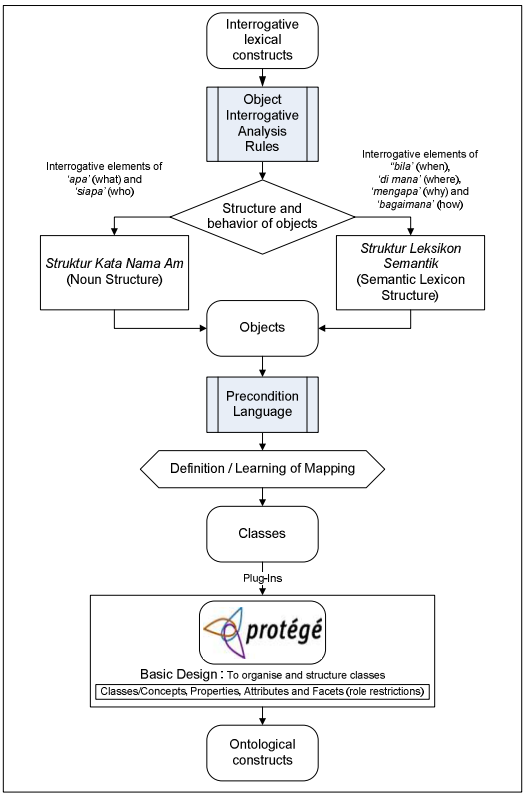

Fig. 1. IKO-Recognizer Process

2.1. Object Interrogative Analysis Rules

Object interrogative analysis rules are capitalizing the Java Object-Oriented Programming (OOP) class encapsulation approach. An encapsulation of class defines the structure and behavior (data and code) that will be shared by a set of objects. Each object of a given class contains the structure and behavior as defined by the class. Most knowledge is made manageable by hierarchical, top-down classification. Java inheritance supports the concept of hierarchical classification, by which one object acquires the properties of another object.

A class is regarded as a logical construct and an object has physical reality. For this, the object interrogative analysis rules use interrogative elements as the most upper class of the object. The structure and behaviors of the objects are implemented through (a) *Struktur Kata Nama Am* (Noun Structure) and (b) *Struktur Leksikon Semantik* (Semantic Lexicon Structure) in order to construct objects.

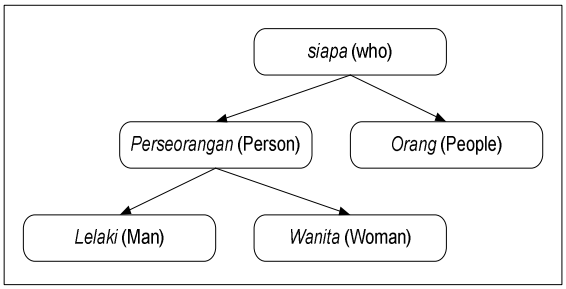

Fig. 2. Example Of 'Siapa' (Who) Object And Its Hierarchical Classification

For the first structure, which is the *Struktur Kata Nama Am* (Noun Structure), the object interrogative analysis rule is defined by combining the structure and behavior of an object with its inheritance and its conceptual modifiers of one or more subclasses in a hierarchical structure. The structure and behaviour of the object are defined by '*kata masuk*' as tagged during the interrogative lexical construct earlier. The '*kata masuk*' for '*penyelidik*' (researcher) is the grammatical information of '*kata nama am*'. It is a noun of '*kata nama am orang*', which refers to as a conceptual of '*Orang*' (People), and has the interrogative element of '*siapa*' (who). Hence, it inherits the general behaviour or properties of its parent '*siapa*' (who). Fig. 2 illustrates the example of object '*siapa*' (who) together with its hierarchical classification.

For the second structure, the *Struktur Leksikon Semantik* (Lexicon Semantic Structure), the object interrogative analysis rule uses the corresponding structure and behaviour of semantic lexicon that defines interrogative elements of '*bila*' (when), '*di mana*' (where), '*mengapa*' (why), and '*bagaimana*' (how). The semantic lexicons of '*bila*' (when) and '*di mana*' (where) correspond to the phrase or proper noun constructed after the semantic lexicon of the interrogative elements. The structure and behaviour of semantic lexicon '*bila*' (when) shows about the time at which an event take place. Whereas, the semantic lexicon of '*di mana*' (where) shows about the place something is in, or is coming from or going to.

Nonetheless, the semantic lexicons of '*mengapa*' (why) and '*bagaimana*' (how) correspond to the predicate after the semantic lexicons of '*mengapa*' (why) and '*bagaimana*' (how). The reason why the semantic lexicons of '*mengapa*' (why) and '*bagaimana*' (how) correspond to the predicate is to describe the meaning of the semantic lexicons and to give information about the sentence. The semantic lexicon of '*mengapa*' (why) talks about the reasons for something which introduces a relative clause after the word reason. Whereas, the semantic lexicon '*bagaimana*' (how) explains the way in which something happens or is done and introduces a statement or fact. The objects of '*mengapa*' (why) and '*bagaimana*' (how) correspond accordingly to their definitions of interrogative element. Their predicates correspond to an attribute of an object defined to support interpretation of the information in unstructured document while maintaining the meaning of the semantic lexicon. The following text is a paragraph taken from [2, 3, 4, 5, 6] a funeral notice for Brian Fielding Frost with Malay language translation.

English sentence: Our beloved Brian Fielding Frost, age 41, passed away Tuesday morning, September 30, 1998, due to injuries sustained in an automobile accident. He was born January 12, 1957 in Salt Lake City.

Malay translation sentence: Brian Fielding Frost yang tercinta, umur 41, meninggal dunia **pagi Selasa, September 30, 1998,** *disebabkan oleh kecederaan dialami dalam satu kemalangan kereta.* Beliau telah dilahirkan **Januari 12, 1957** di Salt Lake City.

The lexicons highlighted in bold are semantic lexicon of '*bila*' (when), italics highlight the semantic lexicon of '*mengapa*' (why), and underlined indicates the semantic lexicon of '*di mana*' (where). The lexicon '*pagi*' (morning) of interrogative element of '*bila*' (when) corresponds to the phrase of date *Selasa*, September 30, 1998. It shows about the time at which things happen. The lexicon '*disebabkan*' of interrogative element of '*mengapa*' (why) corresponds to the predicate of the sentence '*oleh kecederaan dialami dalam satu kemalangan kereta*'. It gives detail about the reason of the death of Brian Fielding Frost. The lexicon '*di*' corresponds with a proper noun of place known as Salt Lake City.

2.2. Precondition Language

The *IKO*-Recognizer also provides a precondition language for defining and learning mapping of the rules, which allows checking certain conditions on these interrogative objects. The structure of the objects is reflected in interrogatively structured form, which represents a generic interrogative ontology. Each concept in the ontology is associated with objects from previous processing as defined by the interrogative analysis rule. For example, building the concept of generic interrogative of 'who' which has successor concepts of 'People', 'Person', 'Woman' and 'Man' is similar with the construction of objects inheritance in the Object-Oriented Programming (OOP).

Correspondingly, the precondition language consists of methods and functions defined for an OOP class. The *IKO*-Recognizer uses the operator CreateConcep to create new class or concept. In current implementation, the supported functions include:

- HasObject: Returns true/false if a certain object corresponds to a specific concept code.
- HasAdjective: Chops off adjective words.
- HasStopWord: Chops off stop words.
- HasVerb: Chops off verb words.
- HasDigit: Chops off digits.
- HasToggle: Chops off toggle words.
- HasUpper: Chops off upper case words.
- HasConcept: Returns new class created.

The *IKO*-Recognizer collectively executes all rules and precondition language. Therefore, when the preconditions are satisfied, the corresponding operator of CreateConcept is activated. This is to create a set of candidate classes, which are automatically generated into a new ontology or are integrated into an existing ontology.

3. RESULTS

IKO-Recognizer facilitates the process of matching and mapping objects from interrogative analysis rule of what/who/when/where/why/how during extraction of the ontological constructs. The *IKO*-Recognizer is implemented as a plug-in to the Protégé knowledge base system. To validate the proposed *IKO*-Recognizer, an experiment is carried out using unstructured documents from a Malay corpus.

The Malay corpus contains 42,733 words from printed materials as well as from articles downloaded from the Internet. Gay and Airasian [26] recommend the use 8% (from 5,000 words) as sample from the population; hence this experiment will use 15% of 42,733 words from the Malay language corpus [27]. The results produced are measured using quantitative retrieval performance of recall and precision metrics [28]. The analysis of results between usage of lexicons and phrases for ontological constructs are shown in Table 1 (where *N* is none existence).

From Table 1, it can be observed that the usage of phrases as interrogative knowledge organization and structuring has increased the average precision as compared to the usage of lexicons. The results above demonstrate that the accuracy of *IKO*-Recognizer is improved on the usage of phrases. The results in the table also signify that the organization of knowledge by phrases performs better than the usage of lexicons in presenting concepts that are more meaningful; i.e., it is able to generate a higher percentage of relevant ontological constructs of expert manual extraction. In other words, the usage of phrases is used by the *IKO*-Recognizer to populate objects where the objects are created as ontology in the Protégé knowledge-base system to capture knowledge from Malay document.

Table I: Comparison On The Usage Of Lexicons And Phrases For Ontological Constructs

	Average Precision	
Recall	Lexicons	Phrases
0.0	N	N
0.1	N	N
0.2	N	N
0.3	0.236111	N
0.4	0.435150	N
0.5	0.634188	N
0.6	0.594406	N
0.7	0.903946	0.783333
0.8	0.917708	0.870833
0.9	0.952634	0.958333
1.0	0.895635	0.974413
Average	0.696222	0.896728
F-Measure	0.704991	0.814257
Percentage	70%	90%

In order to prove that the *IKO*-Recognizer is reliable, a statistical significance test is used to measure the differences between relevant ontological constructs extracted by an expert and the ones generated by the *IKO*-Recognizer. Table 2 and Table 3 show the results of relevant ontological constructs extracted by an expert and the ones generated by the *IKO*-Recognizer, respectively.

Table 2: Number Of Ontological Constructs Extracted By An Expert

*	*'bagaimana'* (how)	*'mengapa'* (why)	*'bila'* (when)	*'di mana'* (where)	*'apa'* (what)	*'siapa'* (who)	#
10	1	1	2	2	14	5	25
11	0	1	3	2	9	7	22
12	1	1	3	2	14	4	25
13	1	1	2	2	18	4	28
14	1	1	2	2	10	4	20
15	1	1	3	2	18	6	31
16	1	1	2	1	11	6	22
17	1	1	2	2	13	4	23
18	1	1	5	2	13	4	26
19	0	1	0	2	6	4	13
20	0	1	1	2	6	1	11
21	1	1	1	2	16	8	29
22	1	0	3	1	8	0	13
@	10	12	29	24	156	57	288

Table 3: Number Of Ontological Constructs Generated By IKO-Recognizer

*	*'bagaimana'* (how)	*'mengapa'* (why)	*'bila'* (when)	*'di mana'* (where)	*'apa'* (what)	*'siapa'* (who)	~
10	1	1	2	2	14	5	25
11	0	1	3	2	9	7	22
12	1	1	3	3	14	4	26
13	1	1	2	2	18	4	28
14	1	1	3	2	10	4	21
15	1	1	3	2	19	6	32
16	1	1	1	1	10	6	20
17	1	1	3	2	11	4	22
18	1	1	4	2	12	4	24
19	1	1	0	2	6	4	14
20	0	1	1	2	7	1	12
21	1	1	2	2	14	8	28
22	0	0	1	1	6	0	8
Σ	10	12	28	25	150	57	282
@	10	12	29	24	156	57	288

Table 4: Differences Between Number Of Constructs By Iko-Recognizer And Expert

*	*'bagaimana'* (how)	*'mengapa'* (why)	*'apa'* (what)	*'bila'* (when)	*'di mana'* (where)	*'siapa'* (who)	>	<	=
10	+	+	+	+	+	+	6	0	0
11	0	+	+	+	+	+	5	0	1
12	+	+	+	+	+	+	6	0	0
13	+	+	+	+	+	+	6	0	0
14	+	+	+	+	+	+	6	0	0
15	+	+	+	+	+	+	6	0	0
16	+	+	-	+	-	+	4	2	0
17	+	+	+	+	-	+	5	1	0
18	+	+	-	+	-	+	4	2	0
19	+	+	0	+	+	+	5	1	0
20	0	+	+	+	+	+	4	1	1
21	+	+	+	+	+	+	6	0	0
22	-	0	-	+	-	0	1	3	2
							64	10	4

Next, the differences between the number of relevant and generated ontological constructs from the Malay unstructured document are observed. Table 4 shows the results where the number of ontological constructs generated by the *IKO*-Recognizer is subtracted from the ones extracted by an expert. From Table 4, it can be observed that the total number of '+' is greater than the total number of '-' and '0'. It is almost 6 times greater than the total number of '-'. It means that there are about 64 constructs out of a total of 78 constructs which accounts for 82%. This implies that the total number of relevant constructs generated by the *IKO*-Recognizer is almost similar with the relevant constructs extracted by an expert. As for the total number of '-', it accounts for 13%, i.e., 10 out of 78 occurrences where the total number of relevant constructs generated by the *IKO*-Recognizer is less than those relevant constructs extracted by an expert, whereas for the total number of '0', it accounts for 5%, i.e., 4 out of 78 occurrences where the total number of relevant constructs generated by the *IKO*-Recognizer and the ones extracted by an expert is equal to zero.

The results in the table show that the *IKO*-Recognizer has the ability as an object recognizer to populate objects and map the objects with ontology engineering to develop ontological constructs which are presented as Malay knowledge representation by concepts. The results also indicate that the *IKO*-Recognizer is able to generate a higher percentage of relevant constructs from an unstructured document. Furthermore, the ontological constructs which are built in the Protégé knowledge-base system are tagged with interrogative contextual information through interrogative lexical constructs and able to hold grammatical information of lexicon entry. This means that the ontological constructs created in the Protégé knowledge-base system are exported into the database management system.

When the sign test is applied in this experiment, the observation is investigated for their significance difference between the number of ontological constructs generated by the *IKO*-Recognizer and the manually extracted constructs by an expert. For each pair, the introduction of the null hypothesis (H_0) and alternative hypothesis (H_1) are required by the sign test which can be generalized as follows:

i. H_0: text in unstructured document cannot be structured to support knowledge organization and structuring, a difference between two types of ontological constructs is *zero* (e.g., the percentage of generated and relevant ontological constructs generated by the *IKO*-Recognizer and the ones manually extracted by an expert is zero); and

ii. H_1: text in unstructured document can have structuring characteristic to support knowledge organization and structuring, a difference between two types of ontological constructs is *positive* (e.g., the percentage of ontological constructs generated by the *IKO*-Recognizer is greater than or equal to the ones manually extracted by an expert).

The results of the sign test on these 13 pairs of observations, using the sign test, are presented in Table 5. From the data in Table 5, it can be observed that the interrogative elements of '*mengapa*' (why), '*siapa*' (who), '*di mana*' (where), '*bagaimana*' (how), and '*apa*' (what) have one-tailed probability when H_0 is true of 0, 0, 0, 0.006, and 0.045, respectively. Since, these values are in the region of rejection for $\alpha = 0.05$, hence the decision is to reject H_0 in favour of H_1. This means that the *IKO*-Recognizer is able to generate significantly better results in populating objects for interrogative elements of '*mengapa*' (why), '*siapa*' (who), '*di mana*' (where), '*bagaimana*' (how), and '*apa*' (what).

Table 5: Frequency Differences Between Iko-Recognizer And Expert (With P Values)

Comparison	*N*	+	-	0	*p*
'mengapa' (why)	12	1	0	12	0
'siapa' (who)	12	1	0	12	0
'di mana' (where)	13	0	0	13	0
bagaimana' (how)	11	2	1	10	0.006
'apa' (what)	13	0	3	10	0.045
'bila' (when)	12	1	3	9	0.073

Unfortunately, this test has shown that there are no significant accuracy differences for interrogative elements of '*bila*' (when). In other words, the knowledge representation of generating ontological constructs by using the *IKO*-Recognizer does not contribute significant accuracy. This can be seen from the value of *p* in Table 5. Since the values of *p* of this interrogative element of '*bila*' (when) is 0.073, the decision is to reject H_1 in favour of H_0.

Results of the experimental testing can be summarized as follows. There is a significant accuracy in populating objects through ontological constructs for interrogative elements of '*mengapa*' (why), '*siapa*' (who), '*di mana*' (where), '*bagaimana*' (how), and '*apa*' (what); and there is no significant accuracy in populating objects through ontological constructs for interrogative element of '*bila*' (when).

In conclusion, the main objective of this experiment is to quantify the accuracy of extracting correctly organized knowledge of populated objects generated through the process of the *IKO*-Recognizer. This experiment proves that the *IKO*-Recognizer has the ability as an object recognizer to populate objects and map the objects with ontology engineering to develop ontological constructs. The results above confirmed that the construction of ontological constructs for interrogative elements of '*mengapa*' (why), '*siapa*' (who), '*di mana*' (where), '*bagaimana*' (how), and '*apa*' (what) have achieved good accuracy results.

The results above also confirmed that the usage of phrases has achieved good accuracy in knowledge organization and structuring. Unfortunately, these accuracy differences are not significant for the interrogative element of '*bila*' (when). This is because the *IKO*-Recognizer is not capable to show a state, condition or continuous activity and particular time, which indicates a period of time.

4. CONCLUSION

Conclusion derived from the development of the *IKO*-Recognizer has pointed out that knowledge extracted from unstructured documents may be successfully represented through knowledge representation of interrogative knowledge organization and structuring. In addition, the interrogative approach through those components have proved that information in Malay unstructured documents can be organized, structured, and transformed in interrogative structured form.

However, there are issues and limitations raised such as the need of familiarization effort on ontology tools in order to develop and structure ontology of Malay language. In current implementation of the interface of IKO-Recognizer is difficult to navigate is not fully automated in creating the corpus. The possibilities for future works is on the usability and functionalities of the *IKO*-Recognizer could be improved by adding token parsing to support homonym and synonym.

REFERENCES

[1] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery: An Overview", In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (eds.), *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, USA: AAAI Press, 1996, pp. 1–34. http://dx.doi.org/10.1609/aimag.v17i3.1230

[2] D.W. Embley, D.M. Campbell, Y.S. Jiang. Y. Ng, R.D. Smith, S.W. Liddle, D.W. Quass, "A Conceptual-Modeling Approach to Extracting Data from the Web", In *Proceedings of the 17th International Conference on Conceptual Modeling*, 1998a, pp. 78-91. DOI: 10.1007/978-3-540-49524-6_7

[3] D.W. Embley, D.M. Campbell, R.D. Smith, S.W. Liddle, "Ontology-based Extraction and Structuring of Information from Data-rich Unstructured Documents", In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, Bethesda MD USA, 1998b. DOI: 10.1145/288627.288641

[4] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, R.D. Smith, "Conceptual-Model-based Data Extraction from Multiple-record Web Pages", *Data and*

Knowledge Engineering, Vol. 31, No. 3, 1999a, pp. 227-251. DOI: 10.1016/S0169-023X(99)00027-0

[5] D.W. Embley, D.M. Campbell, Y.S. Jiang, Y. Ng, "Record-boundary Discovery in Web Documents", ACM SIGMOD Record, *In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, Philadelphia, Pennsylvania, USA, Vol. 28, 1999b, pp. 467-478. DOI: 10.1145/304181.304223

[6] D.W. Embley, "Toward Semantic Understanding – An Approach based on Information Extraction Ontologies", In *Proceedings of the Fifteenth Australasian Database Conference*, Dunedin, New Zealand, 2004, pp. 3-12.

[7] C. Medow, "*Text Information Retrieval System*", Academic Press, San Diego, California, 1992.

[8] E.J. Quigley, A. Debons, "Interrogative Theory of Information and Knowledge", In *Proceedings of the 1999 ACM SIGCPR Conference on Computer Personnel Research*, New Orleans, Louisiana, United States, 1999, pp. 4-10. DOI: 10.1145/299513.299602

[9] T.R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing', *International J. Human Computer Studies*, Vol. 43, No. 5/6, 1995, pp. 907-928. DOI: 10.1006/ijhc.1995.1081

[10] D. Jonassen, K. Beissner, M. Yacci, "*Structural Knowledge: Techniques for Representing, Conveying and Acquiring Structural Knowledge*", Hillsdale (N.J.): Erlbaum, 1993.

[11] M. Gruninger, J. Lee, "Ontology Applications and Design", *Communications of the ACM*, Vol. 45, No. 2, 2002.

[12] C.W. Holsapple, K.D. Joshi, "A Collaborative Approach to Ontology Design", *Communications of the ACM*, Vol. 45, No. 2, 2002.

[13] R. Feldman, I. Dagan, "Knowledge Discovery in Textual Database (KDT)", In *Proceedings of the First International Conference on Knowledge Discovery & Data Mining (KDD-95)* (Montreal, Canada), 1995, pp. 20-21.

[14] R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler, M. Rajman, "Knowledge Management: A Text Mining Approach", In *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM98)*, (Basel, Switzerland), October, 1998, pp. 29-30.

[15] J. Hossein, N. F. M. Sani, L. S. Affendey, I. Ishak, K. A. Kasmiran "Semantic Schema Matching Approaches: A Review", In *Journal of Theoretical & Applied Information Technology*, 62(1), 2014, pp. 139-147.

[16] K. Rajaraman, A.H. Tan, "Mining Semantic Networks for Knowledge Discovery", In *Proceedings of the 3rd IEEE International Conference*, 22 November 2003, pp. 633-636.

[17] X. Lin, D. Soergel, G. Marchionini, "A Self-organizing Semantic Map for Information Retrieval", In *Proceedings of 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1991, pp. 262-269.

[18] S. Kaski, T. Honkela, K. Lagus, T. Kohonen, "WEBSOM – Self-organizing Maps of Document Collections, *Neurocomputing*, Vol. 21, No. 1-3, 1998, pp. 101-117.

[19] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, "Self-organization of a Massive Document Collection, *Neural Networks*, IEEE Transactions, Vol. 11, No. 3, 2000, pp. 574-585.

[20] A. Visa, J. Toivonen, H. Vanharanta H, B. Back, "Knowledge Discovery from Text Documents based on Paragraph Maps", In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[21] A. Kloptchenko, A. Visa, J. Toivonen, H. Vanharanta, "Toward Content-based Retrieval from Scientific Text Corpora", In *Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems* (ICAIS'02), 2002.

[22] K. Lagus, "*Text Mining with the WEBSOM*", PhD Thesis, Helsinki University of Technology, 2002.

[23] K. Lagus, S. Kaski, T. Kohonen, "Mining Massive Document Collections by the WEBSOM Method", *Information Sciences*, Vol. 163, No. 1-3, 2004, pp. 135-156.

[24] A. Saif, M. J. A. Aziz, N. Omar, "Evaluating Knowledge-Based Semantic Measures on Arabic", *International Journal on Communications Antenna and Propagation*, Vol. 4, No. 5, 2014, pp. 180-194.

[25] M. Ribière, P. Charlton, "*Comparison of Ontology Languages: Ontology Overview from Motorola Labs with a Comparison of Ontology Languages*", Retrieved from http://www.fipa.org/ docs/input/f-in-00045/f-in-00045.pdf, 2000.
[26] L.R. Gay, P. Airasian, "*Educational Research: Competencies for Analysis and Application*", Seventh Edition.,Merrill, New Jersey: Upper Saddle River, 2003.
[27] F. Sidi, M.A. Jabar, M.H. Selamat, A.A.A. Ghani, M.N. Sulaiman, S. Baharom, "Malay Interrogative Knowledge Corpus", In American Journal of Economics and Business Administration, 3(1), 171-176. doi:10.3844/ajebasp.2011.171.176
[28] R. Baeza-Yates, B. Ribeiro-Neto, "*Modern Information Retrieval*", New York: Addison Wesley, 1999.