International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 5
ISSN: 2088-5334

Schema Matching for Large-Scale Data Based on Ontology Clustering Method

Harith Alani#1, Saidah Saad#2

#Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
43600 Bangi, Selangor Darul Ehsan, Malaysia
E-mail: 1al_anee76@yahoo.com; 2saidah@ftsm.ukm.my

Abstract*—Holistic schema matching is the process of identifying semantic correspondences among multiple schemas at once. The key challenge behind holistic schema matching lies in selecting an appropriate method that has the ability to maintain effectiveness and efficiency. Effectiveness refers to the quality of matching while efficiency refers to the time and memory consumed within the matching process. Several approaches have been proposed for holistic schema matching. These approaches were mainly dependent on clustering techniques. In fact, clustering aims to group the similar fields within the schemas in multiple groups or clusters. However, fields on schemas contain much complicated semantic relations due to schema level. Ontology which is a hierarchy of taxonomies has the ability to identify semantic correspondences with various levels. Hence, this study aims to propose an ontology-based clustering approach for holistic schema matching. Two datasets have been used from ICQ query interfaces consisting of 40 interfaces, which refer to Airfare and Job. The ontology used in this study has been built using the XBenchMatch which is a benchmark lexicon that contains rich semantic correspondences for the field of schema matching. In order to accommodate the schema matching using the ontology, a rule-based clustering approach is used with multiple distance measures including Dice, Cosine, and Jaccard. The evaluation has been conducted using the common information retrieval metrics; precision, recall, and f-measure. In order to assess the performance of the proposed ontology-based clustering, a comparison among two experiments has been performed. The first experiment aims to conduct the ontology-based clustering approach (i.e. using ontology and rule-based clustering), while the second experiment aims to conduct the traditional clustering approaches without the use of ontology. Results show that the proposed ontology-based clustering approach has outperformed the traditional clustering approaches without ontology by achieving an f-measure of 94% for Airfare and 92% for Job datasets. This emphasizes the strength of ontology in terms of identifying correspondences with semantic level variation.

***Keywords*—automatic schema matching; large-scale data; ontology; clustering; web interfaces.**

I. INTRODUCTION

The vast amount of different data sources has dramatically increased and become accessible through the web interfaces or the so-called deep web [1]. These interfaces have a significant impact on the e-business search engine in which multiple web data sources are unified in a mediated schema. This enables the user to compare among the products [2]. Hence, schema matching has been proposed in order to perform a semantic matching among attributes to find correspondences [3]. Schema matching has attracted many researchers regarding its significance in the data warehousing and integration. However, several challenging issues still facing schema matching due to the various representation mechanism of the data where the attributes can be expressed in different semantic ways.

Many challenging issues have faced schema matching. The common challenging issue is handling large-scale data where several schemas have to be matched concurrency [4]. In fact, handling large-scale data will significantly influence the effectiveness and efficiency. Effectiveness indicates the quality of matching results (i.e. the correctness). Whereas, efficiency indicates the time and resource consumption during the matching process. In order to handle large-scale matching, reducing the search space has been proposed as an effective approach for improving both effectiveness and efficiency. Such approach aims at dividing the data into smaller portions which facilitate the process of finding corresponding among attributes.

Clustering technique has been used as a search space reduction method regarding its ability to partition the data into similar groups. Recently, multiple clustering-based approaches have been proposed for holistic schema matching using different clustering techniques such as k-means and hierarchical. K-means is a rapid clustering technique, but it requires the user to specify the numbers of k clusters which is not easy to obtained in the manner of holistic schema matching [5]. Whereas, hierarchical produces effective results, but restricted due to its time complexity [6]. Additionally, some approaches have been integrated different clustering techniques in order to get better results [7, 8].

However, there is still room for improvement in terms of accuracy. Basically, the traditional clustering approaches have some drawbacks regarding the levels of categories, for instance, the attribute “Date from” could be mutual in the same cluster with the attribute “Date to” which affects the effectiveness of clustering results. This study aims to propose a holistic schema matching method based on ontology clustering.

Several research efforts have been proposed for holistic schema matching. The earliest effort has been presented by Wu et al. [9] where an interactive clustering method has been proposed based on hierarchical clustering for matching web interfaces. It is an efficient method that has the ability to reduce the search space. It achieved an 82% of accuracy for Airfare dataset. In addition, Wu et al. [10] have proposed a merging clustering approach for holistic schema matching where multiple clustering methods have been merged. The experimental results have shown an 85% of accuracy by applying the proposed method on the Airfare dataset. Furthermore, Alofairi & Ahmad [7] proposed an integrated clustering algorithm of k-means and agglomerative hierarchical clustering for holistic schema matching. It exploited name, label and data type with a domain-specific dictionary. It achieves an 89% of accuracy for Airfare dataset. Similarly, Alshaikhdeeb & Ahmad [8] has proposed an integration of correlation clustering and agglomerative hierarchical clustering for holistic schema matching. The proposed method has been carried out on the Airfare dataset. In addition, it obtains a 90% of accuracy.

Ding & Sun [11] have proposed a novel schema matching clustering approach. Such approach has motivated by the attribute position. The authors’ assumption lies on the importance of position where similar attributes from different sources could be matched based on their position. Such approach consists of three phases; first by creating a matrix to represent the statistical occurrence of the attribute position. Whereas, the second phase aims to carry out multiple similarity measures to identify the similarity within the attributes. Finally, the third phase aims to apply a traditional mapping algorithm to group the similar attributes.

The latter approaches proposed for holistic schema matching were relying on clustering-based techniques. Basically, the traditional clustering approaches have some drawbacks regarding the levels of categories

This means that the semantic similarity levels in the clustering approaches are being treated unitedly. In fact, there are different and multiple semantic levels could be countered in the process of schema matching in which the relations such as ‘part-of’ commonly occurs within the schemas (e.g. Date and Date from). In this manner, utilizing a linguistic resource would be very useful in terms of identifying these semantic relations. For this purpose, this paper aims to propose a holistic schema matching method based on ontology clustering.

Utilizing an ontology to perform the schema matching would have a significant impact on the effectiveness where the instances that are incorrectly classified in some clusters using the traditional clustering-based approach, will be handled within the ontology by identifying its semantic level.

II. Materials and Methods

Holistic schema matching is the process of extracting correspondences among large-scale web interfaces repository [8]. The key challenge behind any schema matching approach lies on its effectiveness of matching results, especially when handling large-scale data. In this vein, this study aims to propose an ontology-based clustering approach. This section describes the application of the objective of this study in which the ontology creation, schema matching, and the evaluation is being discussed. As shown in Figure 1, the research methodology consists of two main phases which are (i) ontology creation and (ii) schema matching. These two phases are being illustrated in the following sub-sections.

As shown in Figure 1, every web interface is being represented as a schema which contains multiple fields F_{xi} where i represents the number of fields contained in the schema.

A. *Ontology Creation*

This phase aims to create the ontology by creating the main classes, sub-classes and populating the remaining instances. In order to do so, multiple tasks are being conducted. The following sub-sections are discussing these tasks.

1) Dataset

The datasets used in the experiment composed of two main web interfaces that have been brought from ICQ query namely Airfare and Job [12]. Each dataset contains twenty web interfaces schemas and each schema is being interpreted textually by identifying every single field included with its name and label. Every single schema has been described as a text file that includes the strings of the attribute’s names and labels for each field. Figure 2 depicts a sample of the dataset.

As shown in Figure 2, the sample web interface schema consists of three fields which are ‘Locations’, ‘Job Categories’ and ‘Key Words’. Every field has three main details including Label, Name, and Attribute. Label refers to the text associated with the field in the web interface view (e.g. Locations). Whereas, Name refers to the name that has been given to the field in the database record (e.g. States). Finally, Attribute refers to the data type of the entry for certain fields whether string, integer or date.

2) Pre-processing

This phase aims to get rid of the irrelevant, unnecessary and unwanted information. Obviously, eliminating such information would facilitate and enhance the processing. Multiple tasks have been carried out in order to achieve such phase. These tasks are composed of (i) expanding CamelCase (e.g. turning “CabinClass” into “cabin class”), (ii) removing punctuation (e.g. $^%&#), (iii) removing digits (e.g. 0-9), and (iv) remove stopwords (e.g. of, in, is, etc.).

3) Creating main classes

This task aims to create the main classes of the proposed ontology based on the frequent terms. In fact, these frequent terms may indicate valuable and significant fields. Therefore, TF-IDF has been used to calculate the most frequent terms. TF-IDF has been used in schema matching widely by generating the most frequent terms. In order to calculate the TF-IDF, the term frequency TF has to be computed first as follows:

$$W_s\,(t) = \; TD(t, s\,) \tag{1}$$

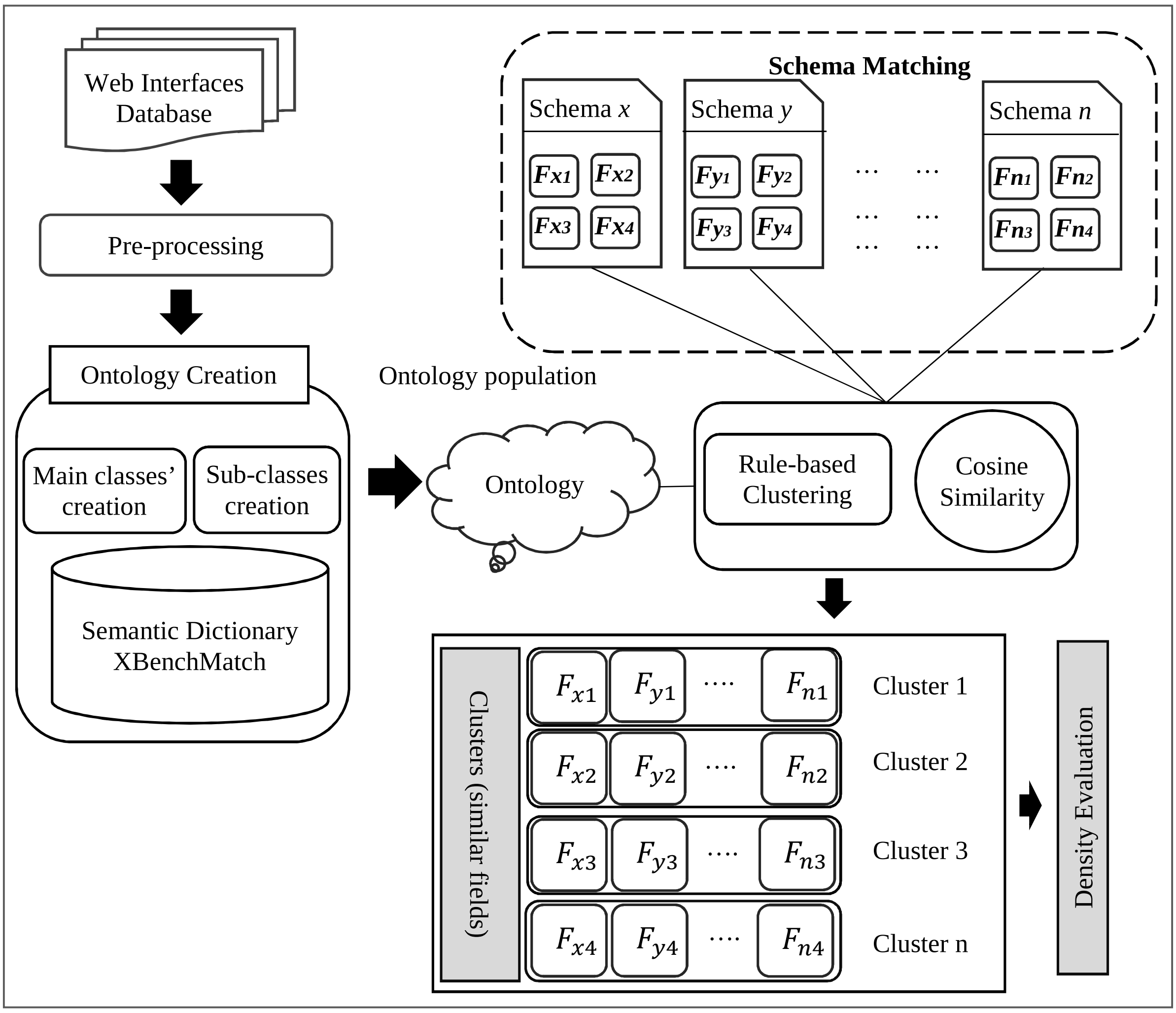


Fig. 1. The framework of the proposed method

where TD (t, d) is the frequency of the term t in the schemas. Then, IDF which aims to provide high weight for rare conditions and low values for common conditions will be computed as follows:

$$IDF_t \; = log\left(\frac{N}{N_t}\right) \tag{2}$$

where N is the number of schemas in the dataset and N_t is the number of schemas that contain the term. Finally, the combination of the two previous methods TF and IDF will be computed as follows:

$$W_t = TF(t, d).IDF_t \tag{3}$$

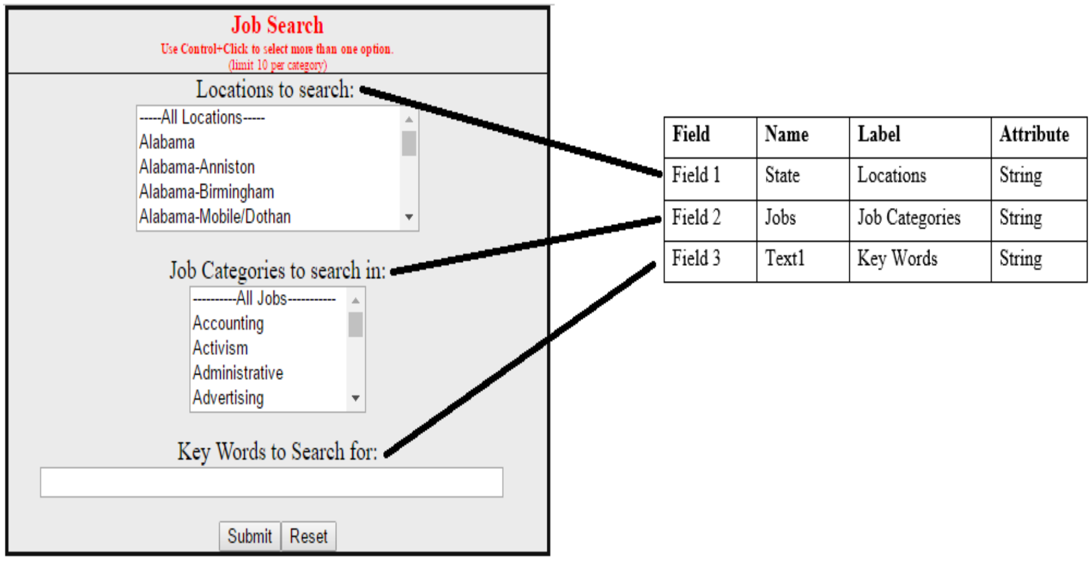
Basically, after applying TF-IDF on all the terms in the schema's dataset, every word will be assigned a value of TF-IDF. This study aims to apply the top-N approach which has been introduced by [13]. This approach aims to select the top candidates with the best values of TF-IDF. In this study, the frequency has been computed for all the schemas in which the top candidates have been selected as 10 best candidates.

4) Creating sub- classes

In order to create the sub-classes, the co-occurrence terms should be declared. Since the sub-classes are belonging to one of the main classes such as 'departure to' which consists of 'departure' and 'to' belongs to the main class 'departure'. Therefore, the co-occurrence would facilitate this process. For this manner, the noun compounds extraction has been utilized. Noun compound is the combination of two or more nouns (e.g. passenger time) or noun with an adjective (e.g. arrival time). In order to identify these noun compounds, it is necessary to utilize a part-of-speech tagging which aims to determine the syntactic class of each word whether noun, adjective or verb. This can be conducted using the following formula:

$$N_0'sN_1 \tag{4}$$

where N1 is a sub-concept of N0. For example, the word ‘city’ in ‘arrival city’ is a sub-concept of ‘arrival’. Table 1 shows a sample of top candidates that have been extracted.



Job Search

Use Control+Click to select more than one option.
(limit 10 per category)

Locations to search:

-----All Locations-----
Alabama
Alabama-Anniston
Alabama-Birmingham
Alabama-Mobile/Dothan

Job Categories to search in:

----------All Jobs----------
Accounting
Activism
Administrative
Advertising

Key Words to Search for:

Field	Name	Label	Attribute
Field 1	State	Locations	String
Field 2	Jobs	Job Categories	String
Field 3	Text1	Key Words	String

Fig.. 2. Sample of the dataset

TABLE I
TOP CANDIDATES OF NOUN COMPOUNDS

Top Candidates	Pattern
Departure city	N + N
Departure month	N + N
Departure day	N + N
Destination date	N + N
Arrival time	JJ + NN
Passenger number	N + N
Departure to	NN + DT

5) Ontology population

After creating the main and sub-classes of the ontology, the remaining instances should be included in these classes. Hence, every instance is being measured in terms of lexical and semantic for determining the appropriate class to be joined. The following tasks describe this process.

N-gram similarity: N-gram similarity aims to identify approximate similarity among two words. In fact, there are numerous words that would share same prefix or suffix, for example, the words ‘dep’, ‘departure’ and ‘departures’ are sharing the same prefix ‘dep’. Such similarity aims to address the approximate match among words. Hence, the N-gram similarity between two words W_1 and W_2 can be stated as follows:

$$Ngram\ (W_1, W_2) = \frac{2C}{A+B} \tag{5}$$

Semantic similarity: in order to identify the similarity between the instances and classes based on the meaning rather than the morphology, semantic dictionary for synonyms is being used. This dictionary is called ‘XBenchMatch’ [14] which has been illustrated using XML and contains rich semantic correspondences for different domain of interest such as Airfare, Job, Computer, Properties and others. Figure 3 depicts a sample of such dictionary.

```
<XS:element name=”Departure”>
   <XS:element synonym1=”Leaving”>
   <XS:element synonym2=”From”>

<XS:element name=”Arrival”>
   <XS:element synonym1=”Destination”>
   <XS:element synonym2=”To”>

<XS:element name=”Job”>
   <XS:element synonym1=”Vacancy”>
   <XS:element synonym2=”Opportunity”>
```

Fig. 3. Sample of XBenchMatch dictionary

The significance demand for using such dictionary lies on the limitation behind the lexical similarity in which the words that are semantically similar but not lexically would not be identified. For example, the two words ‘Job’ and ‘Career’ have the same meaning however, they are lexically different. In the same manner, adding two independent classes that semantically have the same meaning, would be an inefficient way to build the ontology in which the misleading instances would increase the error rate.

Table 2 shows a sample of semantically matches words in both 'Job' and 'Airfare' datasets.

TABLE II
SAMPLE OF SEMANTICALLY RELATED TERMS

Word	Synonym
Departure	Leaving
Departure	From
Arrival	Destination
Arrival	To
Curriculum	Resume
Job	Career
Passenger	infant

B. *Schema Matching*

This phase aims to accommodate the schema matching between the fields of the input schemas. The similar fields are being joined in multiple clusters. Two sub-tasks are being conducted in this phase which stated as follows:

1) Rule-based Clustering

In this study, a rule-based clustering approach has been used to accommodate the schema matching. Such rule-based clustering has been brought from the study of Williams et al. [15] in which a set of rules have been developed to carry out the clustering process. In our study, the rules have been developed based on the taxonomies of the built ontology where the input schema's fields are being processed to identify its branch from the ontology. Let 'Destination day' is an input field that required to be processed using the proposed ontology-based clustering. The taxonomy of the ontology will be shown in Figure 4.

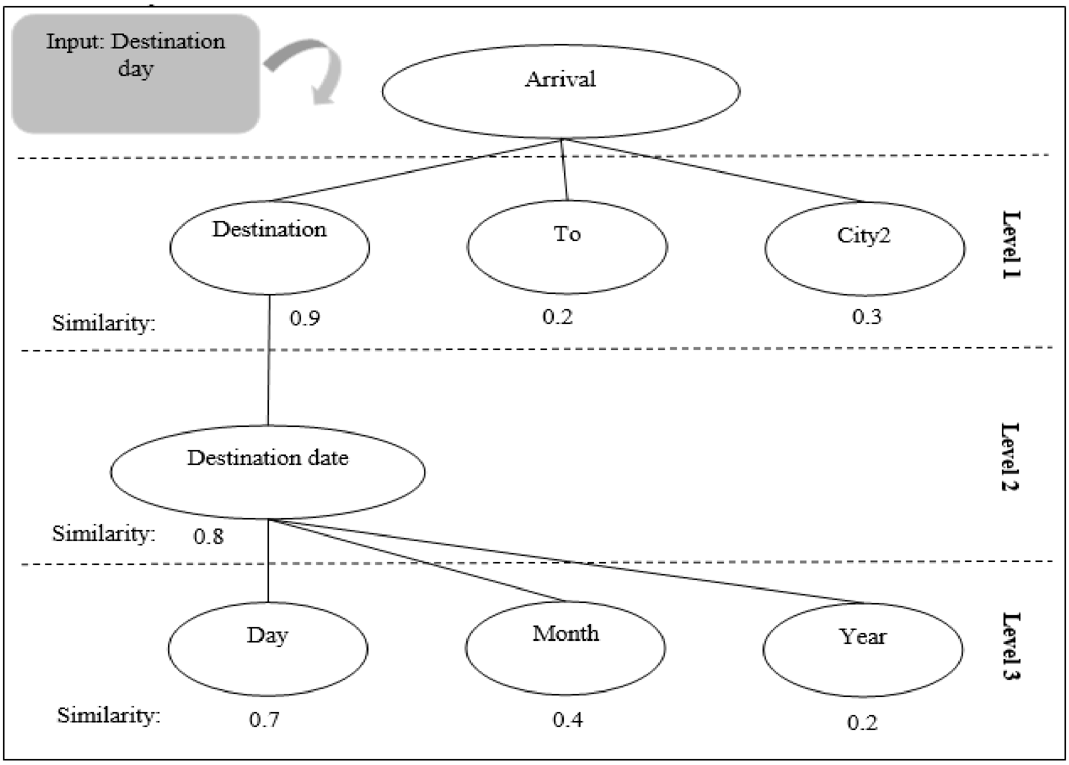


Fig. 4. Taxonomy of the proposed ontology

As shown in Figure 4, the input field 'Destination day' will be processed inside the taxonomy of the ontology. The proposed ontology clustering will identify the similarity between the input and every node in Level 1 including 'Destination', 'To' and 'City2'. With a threshold of similarity as θ = 0.6, the proposed clustering approach will identify which node is the most similar to the input. Note that, the threshold value has been brought from the study of Alshaikheeb & Ahmad (2015). As shown in the figure, the highest value of similarity has been obtained by 'Destination' therefore, the analysis will go further into the containing nodes in Level 2. Level 2 shown one node which is 'Destination date', by measuring the similarity with the input 'Destination day', a similarity value of 0.8 has been obtained. Since the similarity value is above mentioned threshold thus, the analysis will go further into the containing nodes in Level 3. Level 3 shown three nodes including 'day', 'month' and 'year' with similarity values of 0.7, 0.4, 0.2 respectively. Obviously, the highest value was obtained by 'day'. Since there are no further down levels, the proposed method will create a cluster in this level and store the input in such cluster. The pseudo code the proposed clustering method can be depicted in Table 3.

2) Schema Matching

In order to identify the similarity between the input field and the nodes in the ontology, there is a significant demand to use a similarity measure that has the ability to determine the morphological similarity between two strings. For this purpose, the proposed method has utilized three similarity functions including Cosine, Dice Coefficient, and Jaccard Coefficient. These functions can be illustrated as follows:

Cosine

Cosine similarity [16] has been used to identify the similarity between the input field and the nodes from the ontology. The similarity considers the lexical similarity between the input field and nodes, for example, identifying the similarity between 'departure' and 'departures'. Let two strings are $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$, the cosine similarity between them is:

$$SIM_c\left(\overrightarrow{t_a},\overrightarrow{t_b}\right) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{\left|\overrightarrow{t_a}\right| \times \left|\overrightarrow{t_b}\right|} \qquad (6)$$

where $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ are m-dimensional vectors over the term set $T = \{t_1, \ldots., t_m\}$. The results of cosine will be non-negative and ranged in [0,1].

Dice

Similar to Cosine, Dice aims at identifying the lexical similarity between two strings. Let A and B are two strings; Dice can be computed as:

$$D(A,B) = \frac{2|A\,.\,B|}{|A|^2 + |B|^2} \qquad (7)$$

TABLE 1
THE PROPOSED ONTOLOGY-BASED CLUSTERING ALGORITHM

```
Input:   Field $F_i$
         Similarity function $f_x$
         Threshold $\theta$
Output:  Cluster $C$
// Steps:
Compute $f_x$ between the parent nodes and $F_i$
IF (result < $\theta$) Then create Cluster $C$ and store $F_i$
Else {
  Go further down for the child nodes of the maximum parent node in terms of $\theta$
  Compute $f_x$ between the child nodes and $F_i$
  IF (result < $\theta$) Then create Cluster $C$ and store $F_i$
  Else {
    Go further down for the sub-child nodes of the maximum child node in terms of $\theta$
    IF (Max-Child does not contain sub-childe nodes) Then create Cluster $C$ and store $F_i$
    Else {
      Compute $f_x$ between the sub-child nodes and $F_i$
      Repeat Until no further down sub-child found
    }
  }
}
```

Jaccard

Similar to Cosine and Dice, Jaccard aims to determine the morphological similarity among two strings by vectorising them [17]. Let two strings are A and B, Jaccard can be computed as:

$$J(A,B) = \frac{A\,.\,B}{|A|^2 + |B|^2 - A\,.\,B} \qquad (8)$$

C. *Density-based Evaluation*

Prior to discussing the density-based approach, it is necessary to highlight the evaluation method that has been used in this paper to assess the proposed method. Similar to the related work, precision, recall, and f-measure metrics have been used to measure the effectiveness of the retrieval [7, 8]. Precision can be computed as:

$$Precision = \frac{TP}{TP + FP} \qquad (9)$$

where TP is the number of instances that have been retrieved and, at the same time, related to the cluster, and FP is the number of instances that have been retrieved but not related to the cluster. On the other hand, recall can be computed as:

$$Recall = \frac{TP}{TP + TN} \qquad (10)$$

where TN is the number of instances that have not been retrieved even though they are related to the cluster. Finally, f-measure can be computed as:

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (11)$$

Evaluating the performance of a clustering algorithm is a challenging and arguing among the researchers' community [18]. However, one of the common mechanisms for evaluating any clustering algorithm is the density-based approach. The density-based approach aims to assess each cluster independently in which the components of a particular cluster will be examined. Such examination is intended to identify the dominant element. This dominant element will be assigned as the class label of the cluster. In this vein, the dominant element will be determined based on the majority. Apparently, all the instances that are not belonging to the majority will be considered to be incorrectly clustered items.

III. RESULTS AND DISCUSSION

Since the Agglomerative Hierarchical Clustering (AHC) has been demonstrated the best performance in the literature [7, 8]. Therefore, a comparison will be made to compare the performance of the proposed method with AHC using two datasets namely 'Airfare' and 'Job'. Table 4 depicts the 'Airfare' results, while Table 5 depicts the results of 'Job' dataset.

TABLE IV
RESULTS OF AIRFARE

Clustering Approach	Similarity function	Precision	Recall	F-measure
Agglomerative Hierarchical	Cosine	0.80	0.75	0.77
	Dice	0.72	0.64	0.67
	Jaccard	0.67	0.60	0.63
Proposed Ontology-based	Cosine	0.95	0.90	0.92
	Dice	0.90	0.87	0.88
	Jaccard	0.88	0.85	0.86

TABLE V
RESULTS OF JOB

Clustering Approach	Similarity function	Precision	Recall	F-measure
Agglomerative Hierarchical	Cosine	0.80	0.75	0.77
	Dice	0.72	0.64	0.67
	Jaccard	0.67	0.60	0.63
Proposed Ontology-based	Cosine	0.95	0.90	0.92
	Dice	0.90	0.87	0.88
	Jaccard	0.88	0.85	0.86

As shown in Table 4, in terms of the similarity function, Cosine has obtained the highest values of precision, recall, and f-measure using AHC by achieving 85%, 79% and 81%. Similarly, for the proposed ontology-based clustering, Cosine also has shown the greatest results of precision, recall, and f-measure by achieving 97%, 92% and 94%. This superiority of Cosine compared to the other similarity function due to the vectorising mechanism included in its equation which determines the similarity between the strings effectively [17]. Apart from the similarity measures, the greatest values of precision, recall, and f-measure achieved by the traditional clustering approach without ontology (AHC) were 85%, 79% and 81% respectively. In contrast, the greatest values of precision, recall, and f-measure achieved by the proposed ontology-based clustering approach were 97%, 92% and 94% of respectively. Apparently, the proposed clustering approach has outperformed the traditional one significantly.

For the results of 'Job' dataset shown in table 4, the greatest values of precision, recall, and f-measure achieved by the traditional clustering approach without ontology (AHC) were 80%, 75% and 72% respectively. In contrast, the greatest values of precision, recall, and f-measure achieved by the proposed ontology-based clustering approach were 95%, 90% and 92% of respectively. Obviously, the proposed clustering approach has outperformed the traditional one significantly.

As the results revealed, the proposed ontology-based clustering approach has the ability to produce superior results compared to the traditional clustering approach such as AHC. Additionally, it is necessary to accommodate a comparison against the state of the art. Several studies that address the holistic schema matching such as Wu et al. [9] who used the same dataset specifically Airfare and produced 82% of f-measure. Similarly, Pei et al. [19] have applied a k-means clustering approach on the same dataset (Airfare) and produced 88%. In addition, Alofairi & Ahmad [7] have proposed a combination clustering approach using the same dataset (Airfare) and obtained 89% of f-measure. Finally, Alshaikhdeeb & Ahmad [8] have proposed an integrated clustering approach using Airfare dataset and achieved 90% of f-measure. Comparing these results with the proposed method's results (i.e. 94% for Airfare), it is obvious that the proposed method has shown significant enhancement on the clustering results. This is due to the powerful of ontology in terms of identifying multiple and different levels of semantic correspondences.

IV. CONCLUSION

This paper has proposed an ontology-based clustering approach in order to overcome this limitation. Such proposed approach consists of two main phases; creating the ontology and implementing the ontology-based clustering. In fact, the creation of the ontology has been performed using lexical and semantic methods. Lexical methods include term frequency, noun compound extraction, and n-gram similarity. Whereas, a domain specific dictionary which called XBenchMatch has been used as a semantic knowledge. Once the ontology has been created, a rule-based clustering approach has been implemented. Such approach aims to utilize the lexical taxonomy of the proposed ontology in order to perform the holistic schema matching using three similarity measures including Cosine, Dice, and Jaccard. Two benchmark datasets have been used in the experiments including Airfare and Job. In addition, a traditional clustering approach of Agglomerative Hierarchical Clustering (AHC) has been applied too. The experimental results showed that the proposed ontology-based clustering has outperformed the AHC by obtaining an accuracy of 94% and 92% for Airfare and Job respectively.

In fact, the proposed ontology has been intended to serve specific domain of interest. Therefore, building an open domain ontology for the holistic schema matching has the ability to handle more schemas from different domain of interests in future researches.

ACKNOWLEDGEMENTS

This study is supported and funded by the University Kebangsaan Malaysia (UKM).

REFERENCES

[1] Xin Zhong, Yuchen Fu, Quan Liu, Xinghong Lin, and Zhiming Cui, "A Holistic Approach on Deep Web Schema Matching," in Convergence Information Technology, 2007. International Conference on, 2007, pp. 169-174.doi:10.1109/ICCIT.2007.29.

[2] Hai He, Weiyi Meng, Clement Yu, and Zonghuan Wu, "Wise-integrator: An automatic integrator of web search interfaces for e-commerce," in Proceedings of the 29th international conference on Very large data bases-Volume 29, 2003, pp. 357-368.doi:10.1.1.112.8912.

[3] Wei Chen, Huiling Guo, Fang Zhang, Xiaowei Pu, and Xuefei Liu, "Mining schema matching between heterogeneous databases," in Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on, 2012, pp. 1128-1131.doi:10.1109/CECNet.2012.6201642.

[4] Erhard Rahm, "Towards large-scale schema and ontology matching," in Schema matching and mapping, ed: Springer, 2011, pp. 3-27.

[5] Hila Becker, "A survey of correlation clustering," Advanced Topics in Computational Learning Theory, pp. 1-10, 2005.doi:10.1.1.87.3016.

[6] Adel A Alofairi and Kamsuriah Ahmad, "Search space reduction for holistic schema matching: A review," in 2011 International Conference on Research and Innovation in Information Systems, 2011, pp. 1-6.doi.

[7] Adel A Alofairi and Kamsuriah Ahmad, "An integrated clustering method for holistic schema matching," Journal of Theoretical and Applied Information Technology, vol. 68, pp. 294-301, 2014.

[8] Basel Alshaikhdeeb and Kamsuriah Ahmad, "Integrating correlation clustering and agglomerative hierarchical clustering for holistic schema matching," Journal of Computer Science, vol. 11, p. 484, 2015.

[9] Wensheng Wu, Clement Yu, AnHai Doan, and Weiyi Meng, "An interactive clustering-based approach to integrating source query interfaces on the deep web," in Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 2004, pp. 95-106.doi:10.1145/1007568.1007582.

[10] Wensheng Wu, AnHai Doan, and Clement Yu, "Merging interface schemas on the deep web via clustering aggregation," in Data Mining, Fifth IEEE International Conference on, 2005, p. 4 pp.doi.

[11] Guohui Ding and Tianhe Sun, "Schema matching based on position of attribute in query statement," Knowledge-Based Systems, vol. 75, pp. 41-51, 2// 2015.doi:http://dx.doi.org/10.1016/j.knosys.2014.11.005 http://www.sciencedirect.com/science/article/pii/S0950705114003979.

[12] Kevin Chen-Chuan Chang, Bin He, Chengkai Li, and Zhen Zhang, "The UIUC web integration repository," Computer Science Department, University of Illinois at Urbana-Champaign. http://metaquerier. cs. uiuc. edu/repository, 2003.

[13] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in Proceedings of the fourth ACM conference on Recommender systems, 2010, pp. 39-46.doi.

[14] Fabien Duchateau, Zohra Bellahsene, and Ela Hunt, "XBenchMatch: a benchmark for XML schema matching tools," in Proceedings of the 33rd international conference on Very large data bases, 2007, pp. 1318-1321.doi.

[15] Philicity K Williams, Caio V Soares, and Juan E Gilbert, "A clustering rule based approach for classification problems," International Journal of Data Warehousing and Mining (IJDWM), vol. 8, pp. 1-23, 2012.

[16] Anna Huang, "Similarity measures for text document clustering," in Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49-56.doi: http://scholar.google.com.au/scholar.bib?q=info:enBKVjSSXjQJ:scholar.google.com/&output=citation&hl=en&as_sdt=2000&ct=citation&cd=0.

[17] Vikas Thada and Vivek Jaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm," International Journal of Innovations in Engineering and Technology, 2013.

[18] Alex Rodriguez and Alessandro Laio, "Clustering by fast search and find of density peaks," Science, vol. 344, pp. 1492-1496, 2014.

[19] Jin Pei, Jun Hong, and David Bell, "A novel clustering-based approach to schema matching," in Advances in Information Systems, ed: Springer, 2006, pp. 60-69.