Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2

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

AN INTEGRATED CLUSTERING METHOD FOR HOLISTIC SCHEMA MATCHING

ADEL A. ALOFAIRI, KAMSURIAH AHMAD
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia
E-mail: adamoaadam@yahoo.com , kam@ftsm.ukm.my

ABSTRACT

Schema matching is an interesting research topic which had been paid considerable attention by many researchers in the database community. It aims at identifying semantic correspondences between two schemas. Holistic schema matching was proposed to match many schemas at the same time. As an active research topic in the field of schema integration, holistic schema matching tackles the challenge of matching large scale schema. Matching the complete input schemas may not only lead into taking a long execution time, but also poor quality matching results. Therefore, achieving good performance for the schema matching in a large search space is difficult and challenging process. Recently, a number of methods were proposed to solve this schema matching using popular clustering techniques namely k-means or agglomerative hierarchical clustering techniques. These techniques are usually used to reduce the search space, albeit with some drawbacks. However, the existing methods still can to be improved. In order to improve the matching efficiency and clustering process, this paper aims at finding an effective method for holistic schema matching in terms of reducing the searching space. The combination of clustering techniques was proposed. To achieve the main objective, the methodology includes two phases: pre-processing and clustering. The findings of the study revealed that the matching method proposed in this study reduced the searching space by using an integrated clustering technique that rapidly groups the most correspondences attributes in the same clusters. The results of the study prove that this method is effective and promising in holistic schema matching.

Keywords: *Holistic Schema Matching, Clustering, Search Space Reduction.*

1 INTRODUCTION

Due to the remarkable augmentation of diverse data sources, the integration of data has become vital and challenging [8]. The most problematic elements of the integration process are the divergence in naming and the dissimilarity of data structures. These disparities make the integration process more challenging [12]. Nevertheless, schema matching is one of the issues that have drawn the attention of researchers who mainly deal with databases. Traditionally, small scale integration processes were dealt by identifying pair-wise attributes and due to it's limitation these methods did not yield high quality results and thus cannot scale well [9]. The process of schema matching has become more a prominent aspect in variety of applications such as query processing, e-business, data warehouses and semantic web [5]. Matching is the basic manoeuvre in schema information process which involves in producing a map of semantically corresponding elements from any two schema inputs [9]. Usually the schema matching is done manually with a little support of graphical user interface. Hence, it is unavoidably monotonous, difficult, time consuming and non-error free. There are two types of integrations, small scale integration and large scale integration. Based on the input data the large scale integration problems can be classified into two types; (i) Two large size schemas (with thousands of nodes) as in bio-genetic taxonomies. (ii) A large set of schemas (with hundreds of schemas and thousands of nodes) [10]. The large scale schema matching can't be done manually due to the magnitude of variances in the schema. Eventually this has brought to the introduction of a new type of matching known as holistic matching which is efficient enough to match the large scale schemas [3]. The holistic schema matching which takes all the schemas as input and finds all the match among the schemas has become an attractive topic in schema matching. This type of schema matching has been proposed in recent works [3] in order to take advantage of this new opportunity and tackle the challenge of large scale matching.

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

Clustering is the process of grouping data based on their similarities. It is used to spot matching schemas very quickly. Holistic schema matching problem is a combinatorial problem with an exponential complexity, and clustering works as an intermediate technique for a large scale schema matching to improve the efficiency in matching. It reduces the workload for the mapping but this reduction comes with the cost effectiveness [10]. In particular, achieving both effectiveness and efficiencies are two major challenges for large scale schema matching. To address this issue, recently a number of clustering-based approaches are used. The clustering-based approaches used either the K-means [8] or agglomerative hierarchical clustering techniques were proposed to reduce the search pace in order to improve the efficiency of holistic schema matching [8]. However these methods reduce the effectiveness on the matching result. Hence new methods that balance between effectiveness and efficiencies need to be developed [10]. In this study, an integrated clustering method will propose and evaluate toward the enhancement of the clustering performance and effectiveness based on the improvements in the clustering of the attributes.

2 HOLISTIC SCHEMA MATCHING

Holistic schema matches a lot of schemas simultaneously by taking all the schemas as input and finding all the matching among the input schemas. Generally the traditional schema matching works concentrates on small scale integration by identifying pair-wise feature associations between two schemas. The standard approach for pair-wise schema matching is to compare all the elements of the first schema with the elements of the second schema to determine matching schema elements. Apart from having efficiency problems on large scale schema matching, the existence of large search space makes it difficult to accurately identify matching element pairs. Motivated by the needs of integrating large scale data sources, such as the deep Web [3], a new type of schema matching known as holistic schema matching is proposed to improve discover matching in a large set of schemas.

2.1 Web Query Interfaces Matching

With the growing number of data sources accessible over the Web, the integration of these sources is obviously an important problem. It has been observed that at the back of query interfaces are actually hidden a large number of data sources on the Web. In order to integrate the query interface, it is essential to overcome the heterogeneous semantics problem among query interfaces before querying the data. This problem is the most important step is the interface matching. Interface matching is a process of identifying similar fields over multiple query interfaces. Each interface field represents the attribute in the database. The semantic similarity of two fields is evaluated on the similarity of their properties. Two fields are linguistically similar if they have similar names or labels. The clustering approaches or statistical approaches are used for holistic schema matching.

2.2 Search Space Reduction

Due to large search space, holistic schema matching and particularly in matching all input elements may lead to long execution times and poor quality [5]. The process of schema matching is explained as follows: To decide whether elements *s* in schema *S* matches element *t* in schema *T*. The matches must typically examine all other elements in *T* and to make sure that there is no other element that matches *s* better than *t*. This goal of matching adds substantial cost to the matching process. The difficulty of discovering complex matches in the large space is the infinite search process. iMAP proposed by Dhamankar et al. [4] is a system that often reformulates schema matching as a search in a very large or infinite match space. Algergawy et al. [2] proposed a clustered schema matching approach, which is a technique to improve the efficiency of schema matching by means of clustering. Clustering is used to identify regions in the schema repository, which are likely to include good matching for the smaller schema. Then the schema matcher looks for matching only within these regions. From the previous survey and discussion in this study, it appeared that there are many possible ways to perform partitioning for pair-wise large scale schema matching in order to reduce the search space. However, finding the most effective approaches and proposed new methods and techniques for holistic schema type is still an open research problem. Based on the survey performed, we have found that clustering is the most promising techniques used to improve the holistic schema matching performance [1].

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

3 CLUSTERING FOR HOLISTIC SCHEMA MATCHING

Clustering is a process that divides the data into groups of similar objects. Each group, called a cluster, consists of objects that are similar to one another and dissimilar to objects of other groups. Elements within the same cluster share some common property while the elements from different clusters do not. It represents many data objects by few clusters, and hence, it models data by its clusters. Clustering is used to quickly identify regions in the schema repository, which are likely to comprise good mappings for the smaller schema. The schema matcher then looks for mappings only within these regions or clusters. This reduces the matching workload and improves the efficiency. Holistic schema matching problem is a combinatorial problem with an exponential complexity, and clustering works as an intermediate technique for a large scale schema matching to improve the efficiency in matching. Clustering process partitions the mapping elements into clusters. Consequently, this process reduces the size of the search space for the mappings combiner and improves the efficiency. But this reduction may raise issues on cost effectiveness [10].

3.1 CLUSTERING ALGORITHMS

Data clustering research is under vigorous development**.** Contributing areas of research include data mining, statistics, machine learning, spatial database technology, biology and marketing. Many clustering algorithms exist in the literature [1], [2], [3], [9] and most of the clustering algorithms are built around the three concepts: elements, distance measure, and cluster abstraction. In general, the major clustering methods can be classified into:

i. Partitioning methods

The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are similar, whereas the objects of different clusters are dissimilar [7]. The most well known and commonly used partitioning methods are the k-means and k-medoids.

ii. Hierarchical methods

A hierarchical clustering method works by grouping data objects into a tree of clusters and they can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity. In contrast, k-means and its variants have a time complexity, but are thought to produce inferior clusters. Sometimes k-means and agglomerative hierarchical approaches are combined to get the best of both. Indeed, the search of more flexible clustering technique is the important reason of combining both techniques in order to improve the quality of results and the clustering performance [7].

3.2 Clustering Approaches for Holistic Schema Matching

Holistic schema matching essentially applies data mining techniques to discover matching candidates. Recently, a number of clustering-based approaches for holistic schema matching were proposed to improve the matching process performance. These approaches have achieved a remarkable accuracy. They also reduced the workload for a part of the schema matching process by means of search space and execution time. Wu et.al [13] proposed an interactive clustering-based approach to matching query interfaces that captures the hierarchical nature of interfaces. The approach employed a hierarchical agglomerative clustering algorithm. This approach incorporates user interactions to learn the parameters, handled both simple and complex mappings of fields. The approach is highly effective and reduced the search space. It defined the query interface as a problem of identifying semantically similar fields over different query interfaces and identified that any field of interface has three properties: name(f), label(f), and dom(f). To find the 1:1 mappings over the input schemas, a hierarchical agglomerative clustering algorithm is employed. The experiment exploited the name, label, and the domain information for every field. The semantic similarity of two fields is evaluated on the similarity of their properties.

Pei et al. [8] presented a novel clustering-based approach that achieved high accuracy. This approach matched all attributes at once depending on the similarity criteria that used the k-means algorithm. This approach includes three clustering steps: clustered schemas, clustered attributes in the same schema cluster, and cluster attributes across different schema clusters. Unfortunately, the third step is not clearly expressed and only uses attributes name, data types, and label to measure

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

the similarity between the attributes in different schemas.

To improve the efficiency of holistic schema matching the clustering technique works as an intermediate step into existing schema matching algorithms. The clustering technique will partition the schemas into small groups (clusters) and reduces the overall matching load. In this work they proved that clustering effectively contribute the search space reduction in case of large schema matching and improved efficiency. They conclude that clustering can be balanced between effectiveness and efficiency in matching process by tuning the clustering parameters [1].

4 An Integrated Clustering Method

In recent years the holistic schema matching has drawn much attention, due to its efficiency in exploring the contextual information and scalability. This capacity would be useful to automate the process, but as the holistic schema matching is partly subjective, the full automation is not feasible. Based on the literature review, clustering is the most promising techniques used to improve the holistic schema matching performance. Therefore, the main goal is to propose an effective integrated clustering method for holistic schema matching. The proposed method will reduce the matching search space, the matching execution time and improves the clustering process that balance between efficiency and effectiveness. There are a number of stages that should be traversed to achieve the objective of this research. It consists of two phases pre-processing phase and clustering phase as shown in Figure 1.

4.1 Pre-Processing Phase

The dataset used in this work consists of a number of web interfaces schemas that are collected by utilizing the online directories. The Airfare dataset to be used in this work is chosen from the ICQ Query Interface data sets in the UIUC Web Integration repository [11]. These schemas are represented by a text files that contains the attribute's names and labels. The schema of each interface converted to the text file which consist of the attribute name and attribute label for each field in related schema as strings, hence the work in this research exploits the linguistic information (element-based) and the data type (constraint-based) to do the experiments. As in string processing to enhance the similarity measurement, a number of pre-processing steps should be done before the strings comparison process starts, as in the information retrieval (IR) field. Based on the selected dataset the work in this paper performed the following pre-processing steps prior to the clustering phases.

Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2

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

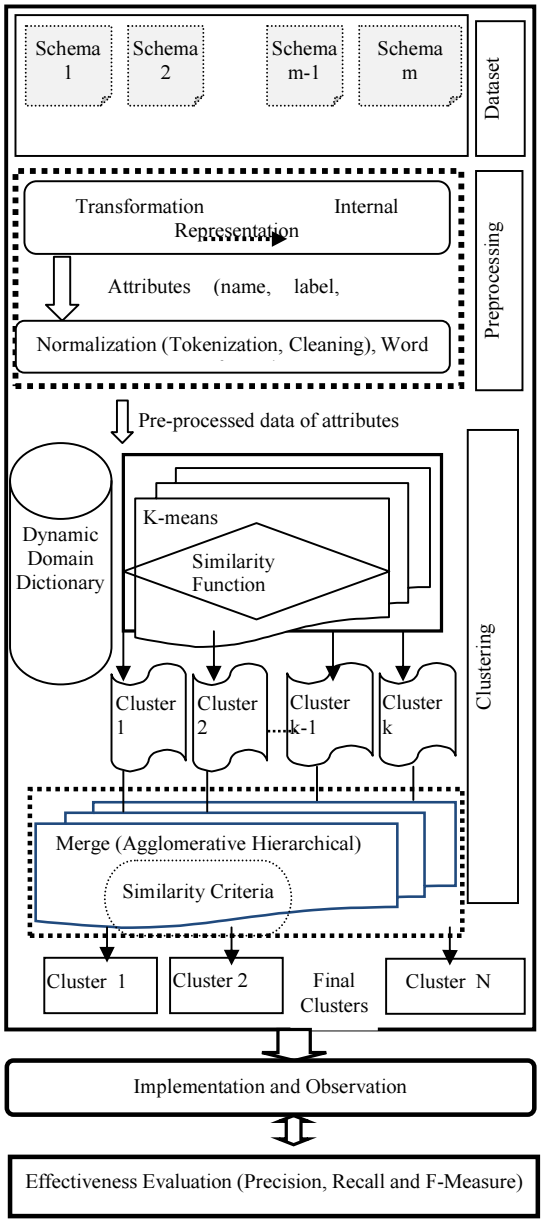

Figure 1: The Proposed Integrated Clustering Method

A. Schemas Transformation

For pre-processing and clustering, the dataset must be converted into the appropriate format.

B. Domain Dictionary

The majority of the modern systems for schema matching use assisting information such as, a synonym dictionary, an abbreviation dictionary, or a specific-domain dictionary to improve the matching results.

C. Data Normalization

The names of attributes frequently contain non-benefited words or characters for matching. Thus, they initially need to be normalized before they are used to compute the similarity of them as strings. For that purpose the following normalizations steps will be applied:

- Tokenization and Cleaning: to deal with concatenated words.
- Stop Word Removing**:** To improve the similarity measure, the unnecessary stop words likes (“or”, “of” , “on”, “in”, “us”) should be eliminated from the input strings
- Word Transformation**:** to expand the abbreviations using the domain dictionary

4.2 Clustering Phase

In schema matching, effectiveness is a term which is concerned with the accuracy and the correctness of matching the results. Whereas efficiency is a term concerned with reducing the execution time and matching space [1]. Recently, a number of methods were proposed to solve this schema matching using popular clustering techniques namely K-means or Agglomerative hierarchical clustering techniques. These techniques are usually used to reduce the search space, albeit with some drawbacks. To improve the matching efficiency and clustering process, the current study aims to combine the k-means or agglomerative hierarchical clustering techniques to produce an effective matching method. This phase integrates two clustering steps: partitioning and merging steps.

4.2.1 THE PARTITIONING STEP

This step represents the first clustering stage. It aims to group the attributes from different schemas that have similarity of names, more than the specified threshold into the same cluster. In this work the input data is divided rapidly by using the k-means local search algorithm as shown in Figure 2.

Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2

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

Input: (D,F_s, T_s)
D. The input dataset
F_s: The similarity function
T_s :The Chosen threshold
Output: *K* clusters
STEPS:
1.Select the attributes of chosen mediated schema as first centroids.
2. *For* each schema S_i in *D* other than the mediated schema *Do*
For each attribute $a \in S_i$ *Do*
Compute the similarity between a and each centroid using F_s.
If no similarity between a and any of the current centroids is > T_s then
Create a new cluster and assign attribute a as the centroid of this cluster
Else
Assign attribute a to cluster c_i whose centroid has the highest similarity with a
If the attributes a, b $\in$ same schema S_i *and* b *in* c_i *Then*
Assign to the cluster c_i the attribute a or b which has the highest similarity with cluster centroid and repeat from step 2.1.1 to compare the other attribute with other clusters.
Re-select the centroids
3.Return the K- clusters of attributes

Figure 2: Clustering algorithm using k-means

The traditional work needs to compare each attribute in each schema, with all the attributes in the other schemas. This procedure increases both the search space and the matching execution time. Hence, the k-means is proposed to find the initial solution (clusters). The clustering process in this work starts by selecting the 14th elements of the first schema as first centroids. The schema which has most of the attributes needs to be matched among all schemas and also has the attribute's names that are similar to other attributes and this called the mediated schema. The system in this work gives the user facility to choose an appropriate threshold value. User can choose different correspondences attributes from different input schemas and measure the similarity between them using similarity measures which are Cosine, Jaccard and Dice [6]. The function for this process is shown in Figure 3. The experiments in this phase exploited the element-based information (names) and the constraint-based (data type) of attribute [8].

The attributes data types classified into the following types: any, integer, string, date, time, year, month, and day.

Input: (a, c , d_t f_s)
a : the attribute
c :the cluster centroid
d : the cluster dataType
f_s :*the function type*
If *dataType(a, c) is not compatible, then return 0*
Case f_s *of*
1: Return *(Cosine(a,c))*
2: Return *(Jaccard(a,c))*
3: Return *(Dice(a,c))*
Output: F_s (a, c)

Figure 3: The string similarity function

4.2.2 The Merging Step

This step represents the second clustering stage. The agglomerative hierarchical clustering technique will be implemented in order to merge the two clusters which has the highest similarity values in each step and have the same data type as shown in Figure 4. This process improves the quality of result by increasing the number of correct matching. The auxiliary information plays an important role to measure the similarity between the attributes. In this phase, the synonyms of names are replaced, using the domain dictionary. The attributes' labels as an additional description will be used to measure the similarity between any two attributes. Each attribute in the dataset is characterized by two properties: name and label. The semantic similarity of two attributes will be evaluated on the similarity of their properties. In this stage the aggregate similarity of two attributes, are calculated based on the linguistic similarity using function in Figure 3. Each of the clusters which resulted from the partitioning phase includes one or more attribute. The combining between the two clustering algorithms in a sequential mode aims to produce the robust clusters and improve the clustering process performance.

5. Experiments and Result

To evaluate the effectiveness of the proposed clustering method, the popular performance measurement in schema matching and information retrieval fields such as precision, recall and F-measure are used. The experiment result of the proposed approach are compared with the result produced by the existing methods such as methods

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

Input: (C_k, C_s, T_s)

C_k : The set of clusters produced from the partitioning phase

C_s : The clusters similarity

T_s : The chosen threshold

Output: (*P*) clusters

STEPS:

Compute the similarity matrix *(M)* for the input clusters.

While there are two clusters in C_k with similarity > T_s

Choose two clusters c_i and c_j, whose similarity is the largest over all pairs of clusters.

Merge c_i and c_j into a new cluster c_v, and remove the clusters c_i and c_j.

Remove all rows and columns associated with c_i and c_j in the matrix (*M*) and add a new row and column for c_v.

Compute similarities of c_v with other clusters using formula (3.2).

Return the (*P*) clusters of attributes.

Figure 4: Agglomerative hierarchical clustering technique

proposed by Pei et al. [8] and Wu et al. [13]. These methods are chosen as comparison because the methods used are similar with the one that proposed in this study. Method proposed by Pei et al. [8] only used k-mean clustering, method proposed by Wu et al. [13] only used hierarchical agglomerative clustering technique, whereas the proposed method combines both of these techniques. The experiment result is shown in Table 1. As shown in the table, the proposed method achieved high similarity values in F-measure and recall compared with techniques proposed by Pei et al. [8] and Wu et al. [13]. It shows that the proposed method improves the schema matching method.

Table 1 Experiment Result

Technique	Precision	Recall	F-measure
[8]	0.99	0.80	0.88
[14]	0.82	0.83	0.82
Proposed method	0.82	0.97	0.89

6. CONCLUSION

For holistic schema matching most of the matching systems that integrates the clustering techniques to the matching process use the clustering techniques to improve the match process performance (efficiency). The main aim of this work is to propose a novel clustering method that reduces the search space for a holistic schema matching into small clusters, which enhances the efficacy and precision. This integrated method was proposed based on the literature review. It blends the widely used and popular clustering techniques. The combination of the k-means and agglomerative hierarchical clustering techniques was implemented and evaluated. The promising outcome proves the quality and effectiveness of the proposed method. In addition this hybrid techniques improves the clustering method performance and gives a new opportunity to solve the existing complexities of matching search space.

REFERENCES

[1] A. A. Alofairi, "An integrated clustering method for holistic schema matching," Master Master thesis, Faculty of information science and technology, Universiti kebangsaan malaysia, Malaysia, 2012.

[2] A. Algergawy, E. Schallehn, and G. Saake. "A schema matching-based approach to XML schema clustering". *In Proceedings of the 10th International Conference on Information Integration and Web-based Applications \& Services*. Austria: ACM. 2008, pp. 131-136.

[3] S. Chuang, and K. Chang.. Integrating web query results: holistic schema matching. *Proceedings of the 17th international conference on Information and Knowledge management (CIKM'08)*, ACM, 2008, pp. 33-42.

[4] R. Dhamankar, Y. Lee, A.H. Doan, A. Halevy, and P. Domingos, "iMAP: discovering complex semantic matches between database schemas". *Proceedings of the International Conference on Management Data SIGMOD, ACM*, 2004, pp. 383-394.

[5] H. Do, and E. Rahm, "Matching large schemas: Approaches and evaluation", *Information Systems journal* 2007, 32(6), pp. 857-885.

[6] J. Han, M. Kamber, and J. Pei, "*Data mining: concepts and technique",*Morgan Kaufmann Publishers, USA. 2011.

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

[7] M. Hadjieleftheriou, and D. Srivastava, "Weighted Set-Based String Similarity". *IEEE Data Eng. Bull.* 2010. 33(1), pp. 25–36 .

[8] J. Pei, J. Hong, and D. Bell, "A Novel Clustering-Based Approach to Schema Matching". *Advances in Information Systems*. In. Yakhno, T. & Neuhold, E. (eds.) 2006, pp. 60-69 Springer Berlin / Heidelberg.

[9] Y. Qian, H. Zhang, J. Song, and Z. Liu, "A new complex schema matching system". *Proceedings of International Conference on International Conference on Innovative Computing and Communication and Asia-Pacific Conference on Information Technology and Ocean Engineering* CICC-ITOE, IEEE, 2010, pp. 292-299.

[10] S. Sellami, A. Benharkat, and Y. Amghar, "Towards a More Scalable Schema Mtching:A Novel Approach". *International Journal of Distributed Systems and Technologies*, 1(1), 2010, pp. 17-39.

[11] http//metaquerier.cs.uiuc.edu/repository/datasets/icq/index.html)

[12] O. Unal, and H. Afsarmanesh, "Semi-automated schema integration with SASMINT", *Knowledge and Information Systems* 23: 2010, pp. 99-128.

[13] W. Wu, C. Yu, A.H. Doan, W. Meng, "An interactive clustering-based approach to integrating source query interfaces on the deep web". *Proceedings of the ACM-SIGMOD International Conference on Management Data* SIGMOD 2004, ACM, pp. 95-106.