Journal of Computer Science

Original Research Paper

Integrating Correlation Clustering and Agglomerative Hierarchical Clustering for Holistic Schema Matching

Basel Alshaikhdeeb and Kamsuriah Ahmad

Faculty of Information Science and Technology, National University of Malaysia, Bangi, Malaysia

Article history
Received: 09-04-2014
Revised: 24-04-2014
Accepted: 18-03-2015

Corresponding Author:
Basel Alshaikhdeeb
Faculty of Information Science and Technology, National University of Malaysia,
Bangi, Malaysia
Email: shaikhdeeb@gmail.com

Abstract: Holistic schema matching is the process of carrying off several number of schemas as an input and outputs the correspondences among them. Treating large number of schemas may consume longer time with poor quality. Therefore, several clustering approaches have been proposed in order to reduce the search space by partitioning the data into smaller portions which can facilitate the matching process. However, there is still a demand for improving the partitioning mechanism by avoiding the random initial solutions (centroids) re-sulted from the clustering process. Such random solutions have a significant impact on the matching results. This study aims to integrate correlation clustering and agglomerative hierarchical clustering toward improving the effectiveness of holistic schema matching. The proposed integrated method avoids the random initial so-lutions and the predefined number of centroids. Several preprocessing steps have been performed with using auxiliary information (domain dictionary). The experiments have been carried out on *Airfare, Auto and Book* datasets from UIUC Web Integration Repository. The proposed method has been compared with K-means and K-medoids clustering methods. As a results the proposed method has outperformed K-means and K-medoids by achieving 0.9, 0.93 and 0.9 of accuracy for *Airfare, Auto and Book* respectively.

Keywords: Schema Integration, Holistic Schema Matching, Correlation Clustering, Agglomerative Hierar-Chical Clustering

Introduction

A huge amount of heterogeneous data sources have expanded and become reachable through the web interfaces or the so-called deep web which refers to the web online accessible databases that dynamically generated in queries of the users (Chang *et al*., 2004). Such interfaces are very crucial for the e-business search engine that provides a unified access to multiple sites, which allow the end-user to search and compare among products easily (He and Chang, 2004). Thus, it is necessary to accommodate heterogeneous semantic between the interfaces queries, this process called schema matching which aims to find the attribute correspondences among the schemas in order to provide a unified interface for the user (Chen *et al*., 2012). Schema matching has become more essential and challenging for many applications such as data warehouses, e-commerce and semantic web (Rahm, 2011). Although it has brought many researchers attentions, schema matching is still an active problem regarding to the variant representation of the data and schema structures (Pei *et al*., 2006). Generally, schema matching is hard because there is often no documentation present with columns that tell us the semantics of the columns, that is, column names and values may be opaque (Jaiswal *et al*., 2013).

Holistic schema matching is the process of carrying off a group of schemas as an input and then outputs the correspondences semantics at the same time (Su *et al*., 2006). Regarding to its efficiency and effectiveness, holistic schema matching has become more challenging and auspicious in terms of solving the problem of large scale matching so that, many approaches have been proposed according to holistic schema matching such as (Chuang and Chang, 2008; Yuchen *et al*., 2009). Furthermore, holistic schema matching has facilitated to attract the attentions of several techniques and approaches in terms of solving large scale schema matching such as clustering technique. Basically, dealing with large scale data may affect the effectiveness of matching results due to the large portions of elements and that may refer to different meanings. As well as, the efficiency may

also effected when matching large scale according to the time and space that would be consumed during the matching task (Rahm, 2011).

Hence, search space reduction has brought the attentions of researchers in order to partition the data into smaller sectors that be easily match with an accurate results. Consequently, clustering has contributed to solve such issue using its features of dividing the data into similar groups. Currently, several clustering-based approaches have been done in terms of solving holistic schema matching using various clustering techniques such as k-means and hierarchical. K-means is a fast clustering technique, but it requires the user to specify the numbers of k clusters which is not easy to attain in the case of holistic schema matching (Wirth, 2010). In contrast, hierarchical produces effective results, but restricted due to its time complexity (Alofairi, 2012). Additionally, some approaches have been integrated both of k-means and hierarchical clustering techniques in order to get better results. However, there is still room for improvement in terms of accuracy toward search space reduction in holistic schema matching. Therfore, this study proposed an integrated method of correlation clustering and agglomerative hierarchical clustering for holistic schema matching. Basically, correlation clustering does not require a user specification and it avoids the random initial solutions (Wirth, 2010). The proposed method utilized the characteristics of both of schema's elements such as columns names and labels and schema's constraints such as the data type of the attributes. Moreover, several preprocessing steps have been performed including transformation, normalization and exploited domain dictionary. The experiments have been carried out on web interfaces which are *Airfare, Auto and Book* datasets from the ICQ Query Interface data sets in the UIUC Web Integration Repository. Eventually, a prototype has been developed based on the proposed method in order to match holistic schemas.

Related Work

He and Chang (2004) presented an integrator tool that match the attribute between query interfaces by exploited names, label and data type. The authors utilized the matching results in order to build a global interface. The matching has been performed using cluster-based method which aims to group the attributes with the same domain into clusters based on their names, then using the semantic (synonyms) such clusters will be merged. The resulted clusters is considered representative attribute for the global interface. Wu *et al*. (2004) proposed an interactive clustering-based approach using agglomerative hierarchical clustering to match query interfaces. It is very effective approach that can reduce the search space and treats simple and complex mappings. Moreover, the approach performs the matching based on the similarity of name, label and domain. Furthermore, Pei *et al*. (2006) proposed a novel clustering-based approach using k-means algorithm. The clustering approach has been done in three steps, (i) clustering schemas, (ii) clustering attributes within the same schema, (iii) clustering attributes in different schema. While, Alofairi (2012) 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. However, the clustering techniques that have been used in the existing approaches have some limitations for instance; k-means requires priori specification of the clusters number and has randomly initial solutions. Therefore, there is still a vital demand for enhancing the clustering techniques in terms of the search space reduction for holistic schema matching.

Materials and Methods

The proposed method consists of three phases. The first phase is preprocessing which contains transformation and normalization.This phase aims to turn the data into an internal representation and eliminating the noisy data. Whereas, the second phase is clustering which contains the integrated correlation clustering and agglomerative hierarchical clustering. Eventually, the third phase which is the evaluation. Figure 1 shows the framework of the proposed method.

Dataset

This research used three datasets each of them contains 20 web interfaces schema's which are collected by utilizing online directories. The chosen datasets are *Airfare, Auto and Book* datasets that brought from the ICQ Query Interface data sets in the UIUC Web Integration Repository (Chang *et al*., 2003). Every single schema has been described as a text file that includes the strings of the attribute's names and labels for each field.

Preprocessing

This phase contains the required procedures that aim to turn the data into a format that can be processed. It includes transformation and normalization which can be illustrated as follows.

Transformation

In terms of pre-processing and clustering, the datasets have to be represented into an appropriate and unified scheme (internal representation) in order to be process-able and executable. Each dataset represented in a table with the following attributes:

(Schema_Number,
Field_Number, Name, Label and do-main)

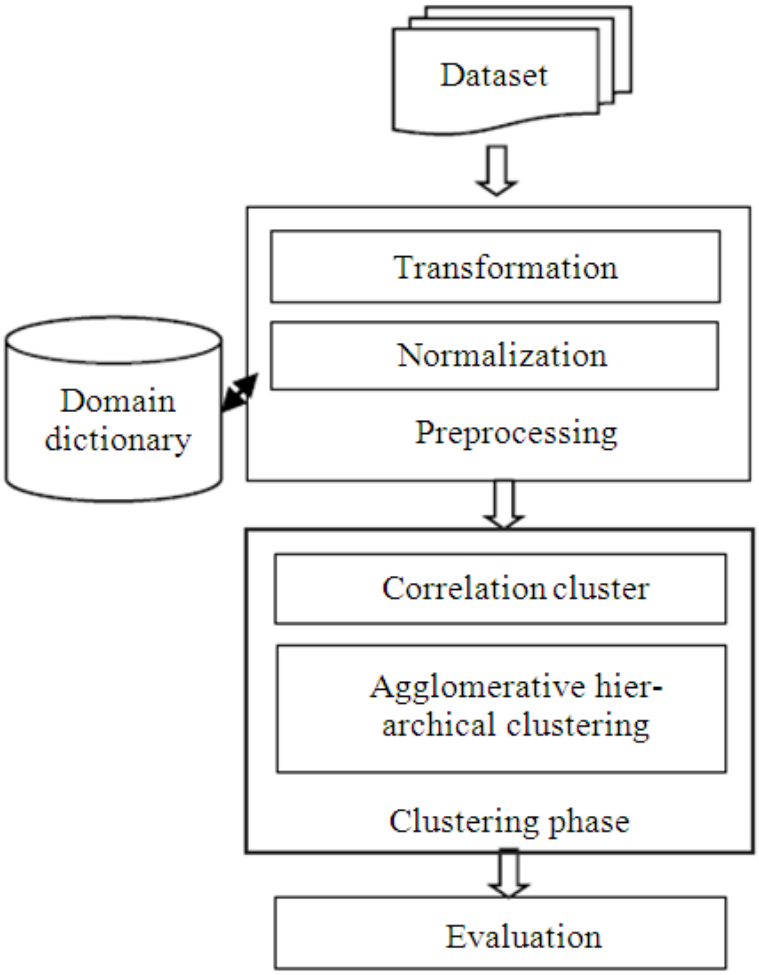

Fig. 1. Framework of the proposed method

Normalization

In order to gain more sophisticated matching the data have to be normalized. Normalization contains several steps which are; (i) Special characters elimination (e.g., &%$#). (ii) Digits elimination (0-9). (iii) Camel-Case splitting (e.g., AuthorName→Author Name). (iv) Stop-words elim-ination (e.g., is, a, an, of). (v) Abbreviations expansion us-ing a domain dictionary for abbreviations (e.g., des→ des-tination). (vi) Synonyms retrieval using a domain diction-ary for synonyms (e.g., class → cabin).

Clustering

The proposed integrated clustering consists of correlation clustering and agglomerative heirarchcial clustering. It aims to assign a point from each schema using correlation clustering. As mention earlier, each schema has several fields thus, in our method point is reffered to a field. The key characteristic of correlation clustering lies on maximizing the agreements (similarities) within a cluster (intra-cluster) and maximizing the disagreements (dissimilarities) between the clusters (inter-cluster) (Wirth, 2010). In order to achieve this objective, Levenshtein Distance have been measured between each field (point) and its neighbor.

Levenshtein Distance LD is the number of procedures have to be performed in order to convert a word to another, those procesdures include insertation, deletion and replacement (Chowdhury *et al*., 2013). LD is usually used to determine the variations among words. Assume x and y are two words, the levenshtein distance between them will be calculated as follow Equation 1:

$$lev_{x,y}(i,j) = \begin{cases} max(i,j) & if\ min(i,j) = 0 \\ min \begin{cases} lev_{x,y}(i-1,j)+1 \\ lev_{x,y}(i,j-1)+1 \\ lev_{x,y}(i-1,j-1)+1_{(xi \neq bj)} \end{cases} \end{cases} \quad (1)$$

where, $1_{(xi \neq bj)}$ is an indicator function equal 0 when $x_i = y_j$ and 1 otherwise. This means that the greater value of LD between two words, the greater dissimilarity among

them. As well as, the lower value of LD between two words, the greater similarity among them.

Hence, the proposed integrated clustering measures the LD values of each field (point) and its neighbor, then it select the group with maximum value of LD and assign them as initial solutions (centroids). Then agglomerative hierarchical clustering has been used to compute the cosine similairty between each field and each centroid. In cosine similarity the two values that wanted to be measure are represented as vectors so that the correlation of those vectors is the Cosine angle between them (Li and Han, 2013). Giving two values $\vec{t}_\alpha$ and $\vec{t}_b$, the cosine similarity between them is Equation 2:

$$SIM_c(\vec{t}_\alpha \vec{t}_b) = \frac{\vec{t}_\alpha . \vec{t}_b}{|\vec{t}_\alpha| \times |\vec{t}_b|} \quad (2)$$

where, $\vec{t}_\alpha$ and $\vec{t}_b$ are m-dimensional vectors over the term set $T = \{t1,\ldots,tm\}$. The results of cosine will be non-negative and ranged in [0,1].

Using a specific threshold the hierarchical clustering merges all fileds with the appropriate centroid. This step is very sensitive due to its influence that associated with the results of matching. Therefore, threshold is very chal-lenging issue that facing off the researchers in the field of schema matching. However, our proposed method per-forms several values of threshold in terms of seeking best matching results. The algorithm of the proposed method is stated below:

```
Generate_Centroids ();
Repeat
  For each schema
    Assign a random field f;
    Compute Lev of f and its neighbors;
     Store schema No. n and its result r in
    list D(n,r);
End For;
Until schema number is reached;
   For each element in D
     Find the maximum r;
     Store all f of n that associated with
     Max r in an array C;
     Assign elements of C as initial
     centroids;
   End For;
Merging ();
   Repeat
      For each fields of schemas
       Compute cosine (Ci,fi)
       If (the result greater than or equal
       the threshold Tc)
          Merge fi into Ci;
       End If;
Until all fields are merged;
```

Results

Basically, three clustering methods have been integrated with hierarchical clustering in terms of portioning which are Correlation clustering, K-means and K- medoids for the three datasets *Airfae, Auto and Book*. The evaluation has been performed using the common information retrieval metrics which are Precision, Recall and F-measure. Eventually, several values of threshold have been adjusted in terms of seeking the optimal. Table 1 shows the results of integrated K-medoids and agglomerative hierarchical clustering.

As shown in Table 1, the results of the integrated k-medoids and agglomerative hierarchical clustering have been described with several values of threshold. It obvious that 0.4 of threshold has achieved the highest values of F-measure which are 0.84, 0.84 and 0.85 for *Airfare, Auto and Book* respectively.

On the same manner, Table 2 shows the results of the integrated K-means and agglomerative hierarchical clustering.

As shown in Table 2, the results of the integrated k-means and agglomerative hierarchical clustering have been described with several values of threshold. It obvious that 0.4 of threshold has achieved the highest values of F-measure which are 0.87, 0.88 and 0.86 for *Airfare, Auto and Book* respectively.

On other hand, Table 3, shows the results of the integrated correlation clustering and agglomerative hierarchical clustering.

As shown in Table 3, the results of the integrated correlation clustering and agglomerative hierarchical clustering have been described with several values of threshold. It obvious that 0.4 of threshold has achieved the highest values of F-measure which are 0.90, 0.93 and 0.90 for *Airfare, Auto and Book* respectively.

Eventually, Table 4 shows the F-measure results of the three clustering method for the three datasets with 0.4 of threshold.

Discussion

As shown in Table 4, correlation clustering has outperformed both of K-medoids and K-means. The outperforming was slightly in both of *Airfare and Book* datasets and remarkable in *Auto* dataset. As expected from other studies such as Velmurugan and Santhanam (2010); Wirth, 2010), unlike k-means and k-medoids, correlation clustering does not require predefined number of clusters and avoid the random initial solution which has a significant impact on the matching results.

Table 1. Results of Integrated K-medoids and agglomera-tive hierarchical clustering

Threshold	Dataset	Precision	Recall	F-measure
0.2	*Airfare*	0.67	0.64	0.65
	Auto	0.68	0.75	0.71
	Book	0.73	0.64	0.68
0.25	*Airfare*	0.73	0.72	0.72
	Auto	0.67	0.69	0.67
	Book	0.58	0.60	0.58
	Airfare	0.65	0.69	0.66
0.3	*Auto*	0.72	0.67	0.69
	Book	0.74	0.72	0.72
0.35	*Airfare*	0.79	0.77	0.77
	Auto	0.81	0.83	0.81
	Book	0.82	0.78	0.79
0.4	*Airfare*	0.84	0.86	0.84
	Auto	0.82	0.88	0.84
	Book	0.87	0.84	0.85

Table 2. Results of integrated K-means and agglomera-tive hierarchical clustering

Threshold	Dataset	Precision	Recall	F-measure
0.2	*Airfare*	0.59	0.55	0.56
	Auto	0.60	0.58	0.58
	Book	0.57	0.59	0.57
0.25	*Airfare*	0.60	0.59	0.59
	Auto	0.62	0.61	0.61
	Book	0.61	0.58	0.59
0.3	*Airfare*	0.67	0.70	0.68
	Auto	0.72	0.73	0.72
	Book	0.75	0.72	0.73
0.35	*Airfare*	0.78	0.76	0.76
	Auto	0.82	0.79	0.80
	Book	0.80	0.79	0.79
0.4	*Airfare*	0.87	0.89	0.87
	Auto	0.90	0.87	0.88
	Book	0.85	0.88	0.86

Table 3. Results of integrated correlation clustering and agglomerative hierarchical clustering

Threshold	Dataset	Precision	Recall	F-measure
0.2	*Airfare*	0.89	0.59	0.71
	Auto	0.82	0.78	0.79
	Book	0.91	0.49	0.63
0.25	*Airfare*	0.86	0.68	0.76
	Auto	0.85	0.80	0.82
	Book	0.88	0.53	0.66
0.3	*Airfare*	0.84	0.72	0.78
	Auto	0.89	0.83	0.85
	Book	0.88	0.52	0.65
0.35	*Airfare*	0.84	0.75	0.79
	Auto	0.90	0.86	0.87
	Book	0.87	0.85	0.86
0.4	*Airfare*	0.87	0.94	0.90
	Auto	0.96	0.93	0.93
	Book	0.90	0.90	0.90

Table 4. F-measure results for the three clustering methods

Method	Airfare	Auto	Book
K-medoids	0.84	0.84	0.85
K-means	0.87	0.88	0.86
Correlation clustering	0.90	0.93	0.90

Conclusion

This research addresses the problem of schema matching by proposing an integrated clustering method consisting of correlation clustering and agglomerative hierarchical clustering toward improving the effectiveness of research space reduction for holistic schema matching. The experiments have been carried out on web interfaces which are *Airfare, Auto and Book* datasets. The proposed method consumed the characteristics of both of schema's elements such as columns names and labels and schema's constraints such as the data type of the attributes. In addition, the cardinality of matching on this research is based on 1:1 matching due to the restriction of the available information that related to the datasets. The proposed method have been evaluated by applying different clustering method and performing a comparison. Correlation clustering has outperformed the other clustering methods.

Funding Information

The authors have no support or funding to report.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Alofairi, A.A., 2012. An integrated clustering method for holistic schemamatching. Faculty of Information Science and Technology.

Chang, K.C.C., B. He, C. Li and Z. Zhang, 2003. The UIUC web integration repository. Computer Science Department, University of Illinois at Urbana-Champaign.

Chang, K.C.C., B. He, C. Li, M. Patel and Z. Zhang, 2004. Structured databases on the web: Observations and implications. SIGMOD Rec., 33: 61-70. DOI: 10.1145/1031570.1031584

Chen, W., H. Guo, F. Zhang, X. Pu and X. Liu, 2012. Mining schema matching between heterogeneous databases. Proceedings of the 2nd International Conference on Consumer Electronics, Communications and Networks, Apr. 21-23, IEEE Xplore Press, Yichang, pp: 1128-1131. DOI: 10.1109/CECNet.2012.6201642

Chowdhury, S.D., U. Bhattacharya and S.K. Parui, 2013. Levenshtein distance metric based holistic handwritten word recognition. Proceedings of the 4th International Workshop on Multilingual, Aug. 24-24, Washington, DC, USA. DOI: 10.1145/2505377.2505378

Chuang, S.L. and K.C.C. Chang, 2008. Integrating web query results: Holistic schema matching. Proceedings of the 17th ACM Conference on Information and Knowledge Management, Oct. 26-30, Napa Valley, CA, USA, pp: 33-42. DOI: 10.1145/1458082.1458090

He, B. and K.C.C. Chang, 2004. A holistic paradigm for large scale schema matching. SIGMOD Rec., 33: 20-25. DOI: 10.1145/1041410.1041414

Jaiswal, A., D.J. Miller and P. Mitra, 2013. Schema matching and embedded value mapping for databases with opaque column names and mixed continuous and discrete-valued data fields. ACM Trans. Database Syst. DOI: 10.1145/2445583.2445585

Li, B. and L. Han, 2013. Distance weighted cosine similarity measure for text classification. Proceedings of the 14th International Conference Intelligent Data Engineering and Automated Learning, Oct. 20-23, Springer, Hefei, China, pp: 611-618. DOI: 10.1007/978-3-642-41278-3_74

Pei, J., J. Hong and D. Bell, 2006. A Novel Clustering-Based Approach to Schema Matching. In: Advances in Information Systems, Springer Berlin Heidelberg, ISBN-10: 978-3-540-46292-7, pp: 60-69.

Rahm, E., 2011. Towards Large-Scale Schema and Ontology Matching. In: Schema Matching and Mapping. Bellahsene, Z., A. Bonifati and E. Rahm (Eds.), Springer Berlin Heidelberg, ISBN-10: 978-3-642-16517-7, pp: 3-27.

Su, W., J. Wang and F. Lochovsky, 2006. Holistic Schema Matching for Web Query Interfaces. In: Advances in Database Technology-EDBT, Ioannidis, Y., M. Scholl, J. Schmidt, F. Matthes and M. Hatzopoulos *et al*. (Eds.), Springer Berlin Heidelberg, ISBN-10: 978-3-540-32960-2, pp: 77-94.

Velmurugan, T. and T. Santhanam, 2010. Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. J. Comput. Sci., 6: 363-368. DOI: 10.3844/jcssp.2010.363.368

Wirth, A., 2010. Correlation Clustering. In: Encyclopedia of Machine Learning, Sammut, C. and G. Webb (Eds.), Springer US, pp: 227-231.

Wu, W., C. Yu, A. Doan and W. Meng, 2004. An interactive clustering-based approach to integrating source query interfaces on the deep Web. Proceedings of the International Conference on Management of Data, Jun. 13-18, ACM New York, pp: 95-106. DOI: 10.1145/1007568.1007582

Yuchen, F., L. Quan, X. Yunlong, Z. Chao and Z. Wenyun *et al*., 2009. Correlated-clustering frame: A Holistic method of deep web schema matching based on data mining. Proceedings of the WRI World Congress on Computer Science and Information Engineering, Mar-Apr. 31-2, IEEE Xplore Press, Los Angeles, CA, pp: 528-533. DOI: 10.1109/CSIE.2009.886