International Journal of Emerging Engineering Research and Technology
Volume 2, Issue 3, June 2014, PP 14-20
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

# Schema Matching with Inter-Attribute Dependencies Using VF2 Approach

**S.Ezhilin Freeda**
Assistant Professor,
Department of Computer Science and Engg,
Sri Ramakrishna Engineering College,
Coimbatore, Tamilnadu, India
*ezhilinfreedasrec@gmail.com*

**T.C.Ezhil Selvan**
Assistant Professor,
Department of Information Technology,
Sri Ramakrishna Institute of Technology,
Coimbatore, Tamilnadu, India
*ezhilselvan85@gmail.com*

**Abstract:** *Schema matching is one of the key challenges in information integration. It is a labor –intensive and time consuming process. The textual similarity of data is to be matched. A two-step technique has been used to address the problem. In the first step, dependencies within tables have been measured and a dependency graph has been constructed. In the second step matching node pairs across the dependency graph have to be identified by running a graph matching algorithm. The proposed system is to be implemented with Mahalonobis distance metric in order to provide better metrics and the graph matching is to be implemented by using VF2 algorithm to improve the accuracy of schema matching. The accuracy of the approach has to be experimentally validated.*

**Keywords:** *Schema matching, Mahalanobis distance metric, VF2 algorithm.*

## 1. INTRODUCTION

The schema-matching problem at the most basic level refers to the problem of mapping schema elements (for example, columns in a relational database schema) in one information repository to corresponding elements in a second repository. Consider a classical schema mapping problems where two employee tables are integrated. One logical approach is to compare the attribute names across the tables. This is termed as schema based matching. Such an approach will not be effective because different institutions may use different names for semantically identical attributes, or use similar names for unrelated attributes. When schema based matching fails, the next approach is to look at the data values stored in the schemas.

This technique is called an instance-based technique [3]. Instance based mapping fails, because of the inability to distinguish different columns over the same data domain and inability to find matching columns using values drawn from different domains. In such cases Instance based approach will fail to identify the correspondence between two columns.

## 2. LITERATURE REVIEW

### 2.1 Mutual Information and Entropy

Mutual information (MI) is an information theoretic similarity measure assessing the dependence of one random variable on another. Intuitively, maximizing MI is equivalent to minimizing the joint entropy relative to the marginal ones as shown in [5]. MI can thus be thought of as a measure of how well one random variable explains the other, i.e. a measure of the amount of information A contains about B and vice versa. The mutual information corresponding to the two random variables is defined as,

$$\mathrm{MI}(X;Y) = \sum_{x \in N} \sum_{y \in S} P(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Entropy:

Let X is an attribute with alphabet N, and p(x) probability distribution of X. The entropy H(X) is defined by

$$P(X)=-\sum p(x)\log p(x) \quad (2)$$

Conditional entropy:

Let X and Y be two attributes with alphabets N and S respectively. Conditional entropy of X and Y can be defined as:

$$H(Y/X)=\sum_{x\in N}\sum_{y\in S} p(x,y)\log p(x/y) \quad (3)$$

### 2.2 Approaches for Schema Matching

Approaches to schema integration can be broadly classified as schema information or schema and instance level information as shown in [7][8]

- *Instance vs schema:* matching approaches can consider instance data (i.e., data contents) or only schema-level information.
- *Element vs structure matching:* match can be performed for individual schema elements, such as attributes, or for combinations of elements, such as complex schema structures.
- *Language vs constraint:* a matcher can use a linguistic based approach (e.g., based on names and textual descriptions of schema elements) or a constraint-based approach (e.g., based on keys and relationships).

## 3. METHODOLOGY

In this project to improve the accuracy of schema matching, it is proposed that the Mahalonobis distance metric will be calculated and the graph matching will be done by using VF2 algorithm.

### 3.1 Mahalonobis Distance Metric

The Distance metric is a key issue in many machines learning algorithm. For example, Kmeans and K-Nearest Neighbor (KNN) classifier need to be supplied a suitable distance metric, through which neighboring data points can be identified. The commonly-used Euclidean distance metric assumes that each feature of the data point is equally important and independent from others.

The Mahalanobis distance is a measure between two data points in the space defined by relevant features. Since it accounts for unequal variances as well as correlations between features, it will adequately evaluate the distance by assigning different weights or importance factors to the features of data points is shown in [9]. Only when the features are uncorrelated, the distance under a Mahalanobis distance metric is identical to that under the Eu- clidean distance metric

### 3.2 VF2 Algorithm

Graph matching is a computationally challenging due to the combinatorial nature of the set of permutations. When graphs are used for the representation of structured objects, then the problem of measuring object similarity turns into the problem of computing the similarity of graphs, which is also known as graph matching

In the existing Hill climbing approach, Entropy and mutual information will be calculated for two different tables and then euclidean distance metric will be found. Based on the attributes, to find a mapping whose Euclidean distance is closest to the optimum for the two tables.

The VF2 algorithm obtains the best performance for graphs of small size and for quite sparse graphs. In case of bounded valence graph, if the valence is small, VF2 is always the best algorithm as shown in fig 1.

The VF2 algorithm can be used for all morphism types as shown in [1]. The fastest variant runs at a much higher speed than the original implementation for the problem of graph isomorphism. Performance Comparision is made for Hill Climbing and Vf2 algorithms

## 4. IMPLEMENTATION

### 4.1 Modeling Dependency Relation

The dependency graphs can be constructed by calculating the pair wise mutual information over all pairs of attributes in a table and structuring them in an undirected labeled graph the label on a node represents the entropy of the attribute, which is equivalent to its mutual information with itself or self-information is shown in Fig 2 & 3. The dependency graph can be modeled in a simple symmetric square matrix of mutual information, which is defined as follows:

Let S be a schema instance with n attributes and $a_i (1 \le i \le n)$ be its i$^{th}$ attribute. The dependency graph of schema S using square matrix M can be defined by,

$$M = (m_{ij}), where\ m_{ij} = MI(a_i; a_j), 1 \le i, j \le n \quad (7)$$

The intuition behind using mutual information as a dependency measure is twofold:

- It is value independent; hence, it can be used in uninterrupted matching.
- It captures complex correlations between two probability distributions in single number, which simplifies the matching task in the second stage of our algorithm.

### 4.2 Hill Climbing

Hill climbing is a simple nondeterministic, iterative improvement algorithm [2]. The HC algorithm is simply a loop that moves, in each state transition, to a state where the most improvement can be achieved. A state represents a permutation that corresponds to a mapping between the two graphs.

The set of all states reachable from one state in a state transition, to a set of all permutations obtained by one swapping of any two nodes in the permutation corresponding to the current stateVF2 Algorithm

### 4.3 VF2 Algorithm

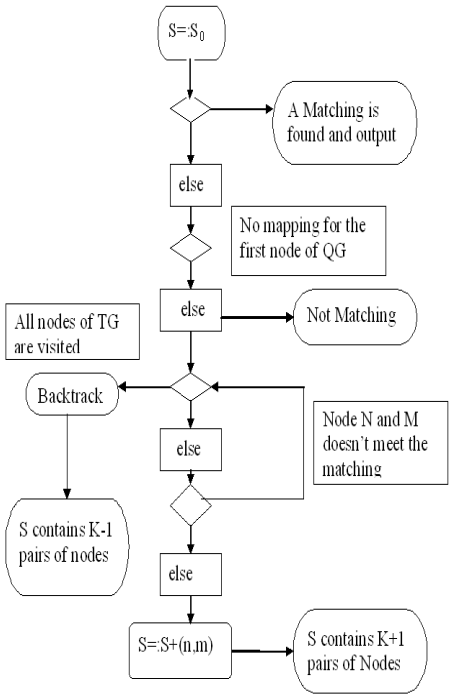


**Fig1.** *Flow Graph of VF2 Algorithm*

The VF2 algorithm explores the search graph by means of a depth-first-search and uses pruning techniques to reduce the size of the generated solution tree. It tries to extend an existing mapping of nodes and edges until a full mapping are reached, starting from the empty mapping.

## 5. RESULTS

### 5.1 Modeling Dependency Relation

An Input table has been considered with ten columns (Fig 2). For each attribute entropy and Mutual information has been calculated and the dependency graph was drawn as shown in Fig 3.

table1 : Table

| studname | gender | rollno | address | degree | Grade | major | age |
|---|---|---|---|---|---|---|---|
| M.Ambiga | Female | 105 | Palladam | MBA | A | Acc. | 27 |
| P.Prakash | Male | 109 | Pollachi | ME | B | AE | 22 |
| A.Suruthi | Female | 117 | Trichy | M.SC | C | Agri | 25 |
| H.Babu | Male | 112 | Trinelveli | MCA | D | App | 24 |
| C. Vijay | Male | 120 | Tripur | M.phil | E | Biology | 23 |
| L.Kathir | Male | 116 | Nagercoil | B.Sc | O | Bio-Tech | 19 |
| P.anitha | Female | 102 | Dharapuram | ME | S | CAD/CAM | 29 |

**Fig 2.** *Input Table*

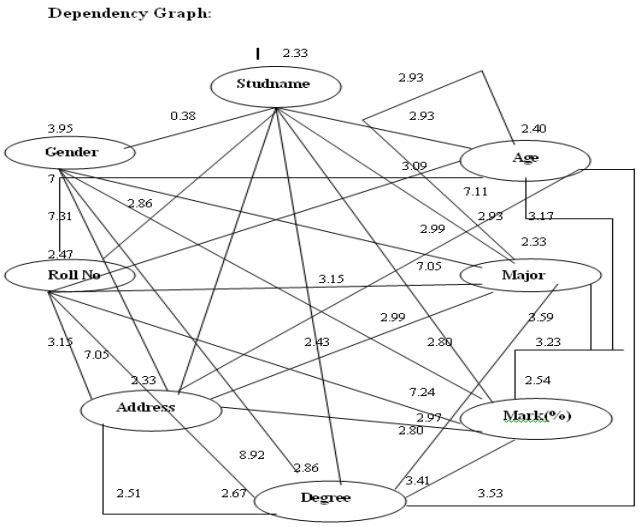


**Fig3.** *Dependency Graph*

### 5.2 Hill Climbing

table2 : Table

| studname | gender | rollno | address | degree | percentage | major | age |
|---|---|---|---|---|---|---|---|
| M.Ambiga | Female | 105 | Palladam | MBA | 100 | Acc. | 27 |
| P.Prakash | Male | 109 | Pollachi | ME | 90 | AE | 22 |
| A.Suruthi | Female | 117 | Trichy | M.SC | 95 | Agri | 25 |
| H.Babu | Male | 112 | Trinelveli | MCA | 80 | App | 24 |
| C. Vijay | Male | 120 | Tripur | M.phil | 85 | Biology | 23 |
| L.Kathir | Male | 116 | Nagercoil | B.Sc | 70 | Bio-Tech | 19 |
| P.anitha | Female | 102 | Dharapuram | ME | 75 | CAD/CAM | 29 |
| E.Deepa | Female | 119 | Mysore | M.D | 60 | Cardio | 38 |

**Fig 4.** *Input Table with Single column change*

In order to compute the similarity between a pair of tables, the original input table (Fig 2) and a modified table with a single column change (Fig 4) has been considered. Mutual information was calculated for the tables and then Euclidean distance metric was computed as shown in table 1. Based on distance the degree of match between two tables can be determined.

**Table1**. *Euclidean Distance between Tables of Figure 2 and 4.*

| Columns | Table 1 MI Calculation | Table 1 MI Calculation | Euclidean Distance |
|---|---|---|---|
| 01 | 0.38123094930796764 | 0.38123094930796764 | 0.0 |
| 02 | 2.8662657447607374 | 2.8662657447607374 | 0.0 |
| 03 | 2.9957322735539833 | 2.9957322735539833 | 0.0 |
| 04 | 2.510355038967773 | 2.510355038967773 | 0.0 |
| 05 | 2.487273562476667 | 2.9957322735539833 | 0.5084587110773162 |
| 06 | 2.9957322735539833 | 2.9957322735539833 | 0.0 |
| 07 | 2.93099900915736 | 2.93099900915736 | 0.0 |
| 12 | 7.3160031087656652 | 7.3160031087656652 | 0.0 |
| 13 | 7.058727718990701 | 7.058727718990701 | 0.0 |
| 14 | 8.921928151371127 | 8.921928151371127 | 0.0 |
| 15 | 9.480865918774338 | 7.058727718990701 | 2.422138199783637 |
| 16 | 7.058727718990701 | 7.058727718990701 | 0.0 |
| 17 | 7.119165169052917 | 7.119165169052917 | 0.0 |
| 23 | 3.1573875605959083 | 3.1573875605959083 | 0.0 |

Total Euclidean Distance: 5.617

In order to show that dissimilar tables have a greater Euclidean distance between them two columns were changed in the original input table and Euclidean distance was computed. The resulting values are shown in table 2.

table2 : Table

| studname | gender | rollno | registerno | degree | percentage | major | age |
|---|---|---|---|---|---|---|---|
| M.Ambiga | Female | 105 | 08101 | MBA | 100 | Acc. | 27 |
| P.Prakash | Male | 109 | 08102 | ME | 90 | AE | 22 |
| A.Suruthi | Female | 117 | 08103 | M.SC | 95 | Agri | 25 |
| H.Babu | Male | 112 | 08104 | MCA | 80 | App | 24 |
| C. Vijay | Male | 120 | 98105 | M.phil | 85 | Biology | 23 |
| L.Kathir | Male | 116 | 08106 | B.Sc | 70 | Bio-Tech | 19 |
| P.anitha | Female | 102 | 08107 | ME | 75 | CAD/CAM | 29 |
| E.Deepa | Female | 119 | 08109 | M.D | 60 | Cardio | 38 |
| R.Pratheeb | Male | 111 | 08110 | MD,DM | 65 | Cardiologists | 40 |
| T.Ramesh | Male | 125 | 08111 | M.Phil | 50 | Chem | 35 |
| T.latha | Female | 103 | 08112 | BDS | 55 | Dental | 21 |

**Fig 5.** *Input Table with Two column Change*

| Columns | Table 1 MI Calculation | Table 1 MI Calculation | Euclidean Distance |
|---|---|---|---|
| 01 | 0.38123094930796764 | 0.38123094930796764 | 0.0 |
| 02 | 2.538017969118249 | 2.8662657447607374 | 0.3282477756424882 |
| 03 | 2.9957322735539833 | 2.9957322735539833 | 0.0 |
| 04 | 2.510355038967773 | 2.510355038967773 | 0.0 |
| 05 | 2.552006826873291 | 2.9957322735539833 | 0.4437254466806921 4 |
| 06 | 2.9957322735539833 | 2.9957322735539833 | 0.0 |
| 07 | 2.93099900915736 | 2.93099900915736 | 0.0 |
| 12 | 9.077345555865058 | 7.3160031087656652 | 1.7613424470994063 |
| 13 | 7.058727718990701 | 7.058727718990701 | 0.0 |
| 14 | 8.921928151371127 | 8.921928151371127 | 0.0 |
| 15 | 8.897034366574001 | 7.058727718990701 | 1.8383066475833 |
| 16 | 7.058727718990701 | 7.058727718990701 | 0.0 |
| 17 | 7.119165169052917 | 7.119165169052917 | 0.0 |

Total Euclidean Distance: 10.640

**Table 2.** *Euclidean Distance between Tables of Figure 2 and 5.*

## 6. DISCUSSION

The implementation of graph matching algorithm is in progress. After completion, comparison will be made between Hill Climbing and VF2 algorithm. VF2 Algorithm is expected to yield better accuracy than Hill Climbing.

## 7. CONCLUSION

Schema-matching techniques are based on the use of data interpretation and structural similarity. A schema matching problem can be reduced to a traditional graph matching problem by capturing hidden dependencies between attributes and structuring them labeled as a graph that takes into account the dependency relations among the attributes.

The proposed system is to be implemented with Mahalonobis distance metric in order to provide better metrics and the graph matching is to be implemented by using VF2 algorithm to improve the accuracy of schema matching.

Comparison will be made for Hill Climbing and VF2. The proposed algorithm VF2 definitely will provide better accuracy than Hill climbing.

## REFERENCES

[1] P.Foggia, C.Sansone, M.Vento, "A performance comparison of five algorithms for graph isomorphism", Journal of the Association for computing machinery vol 10, no.4, 2001.

[2] J.Kang and Jefferey F.Naughton, 'Schema matching using inter-attribute dependencies', IEEE Transactions on Knowledge and Data Engineering., vol 20, no.10, 2008.

[3] J.Kang and J.F.Naughton, "On schema matching with opaque column names and data values", Proceedings.acm sigmod ,2003.

[4] Kilian Q. Weinberger, "Fast solvers and efficient implementations for distance metric learning", IEEE Computer society.org, 2008.

[5] Manolis I.A Lourakis, Antonis A. Argyros and kostas maria, "A graph-based approach to corner matching using mutual information as a local similarity measure", IEEE Computer society.org, 2004.

[6] Paul Johannesson, "Using conceptual graph theory to support schema integration "Proceedings ACM int'l conf. knowledgediscovery and data mining (kdd '04),2004, pp. 148-157

[7] E.Rahm and P.A.Bernstein ," On matching schemas automatically",the vldb j., 2001, vol.10, no 4.

[8] E.Rahm and P.A. Bernstein, "A survey of approaches to automatic schema matching," the vldb j., 2001, vol. 10, no. 4.

[9] Shiming Xiang , Feiping Nie, Changshui Zhang," Learning a Mahalanobis Distance Metric for Data Clustering and Classification", proceeding ACM , 2008, pp-23-26.

## AUTHORS' BIOGRAPHY

**S.Ezhilin Freeda** received the BE degree in Computer Science and Engineering from Anna University, Chennai, India, in 2008 and the ME degree in Computer Science and Engineering from Anna University, Chennai, India, in 2010. Her research interests include wireless networks and Data Mining.

**T.C.Ezhil Selvan** received the BE degree in Computer Science and Engineering from Anna University, Chennai, India, in 2006 and the ME degree in Software Engineering from Anna University, Chennai, India, in 2009.He is currently working towards his PhD degree in Information and Communication Technology discipline at Anna University, Chennai, India. His research interest includes Mobile Computing, Wireless Communication and Data Mining.