ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Study and Comparison of Methods for Fuzzy Data Equivalence

Ankita Srivastava*and Awadhesh Kumar Sharma
Department of Computer Science & Engineering
Madan Mohan Malaviya Engineering College
Gorakhpur-273010, U.P., India
ankitasrivastava1317@gmail.com*
akscse@rediffmail.com

Abstract: In the current scenario database integration plays a vital role in large number of applications. These applications require the data from various databases be integrated to provide a unified view of data. Much work has been done in the field of crisp database integration but comparatively very less work has been done in the field of fuzzy database integration. Comparison of fuzzy data plays a significant role in fuzzy database integration, therefore, this paper provides a brief overview of the some of the existing methods for comparing fuzzy data and comparison of these methods is also done.

Keywords: Fuzzy database, Fuzzy equivalence, Possibility distribution of fuzzy data.

I. INTRODUCTION

In the real world the information available is not always precise i.e. it is often uncertain or imprecise. Therefore, conventional data models have been extended so that imperfect and imprecise information can be stored and manipulated. It is not always possible to store all the information or data in one database; therefore the need of integrating two data sources has arisen. The process of data integration has two parts: schema integration and instance integration [1]. The main problems faced in schema integration are entity type incompatibility and attribute homonym problem. In entity type incompatibility the different entity instances possess properties depicted by the attribute. For example, weight of a person is incompatible with the weight of a machine. In attribute homonym problem, an entity's different properties are depicted by same attribute. For example, phone number an employee in one database may represent office phone number and residence phone number in another. The problems faced in instance integration are entity identification and attribute value conflict. In entity identification the aim is to identify tuples from the source databases which represent same entity in the real world.

The attribute value conflict shows the conflict in value of attribute representing the same property of the real world entity. The attribute value conflict resolution can't be done until entity identification process is done. Techniques for schema integration [2] have been extensively studied but very little research has been done in the field of instance integration. Schema integration has to be performed before instance integration. Entity identification can be easily done in simple databases but is complex for fuzzy database. For entity identification in fuzzy database the fuzzy data need to be compared so that one is able to know whether they represent same entity or not. This paper gives the brief study of various methods for fuzzy data equivalence. If the fuzzy data belong to same relation then by checking the equivalence one can easily detect redundancy in the database so that inconsistency can removed by deleting the duplicate tuples. In view of instance integration of fuzzy databases, comparison fuzzy data which belong to different relations is helpful in identifying whether the fuzzy data being compared represent same entity or not. In the next section fuzzy data equivalence methods have been described. Then these methods have been compared in table 1, and in section 3, problems in these methods has been stated.

II. FUZZY DATA EQUIVALENCE METHODS

This paper lists methods for fuzzy data equivalence in which R is a fuzzy relation and some attributes take imprecise or uncertain values defined by possibility distributions. The methods are divided in two groups [3] depending on how comparison of possibility distribution is done. The two groups of approaches are:

Case 1: In this case, only the possibility degrees of each element in the domain of the given attribute are compared.

Case 2: In this the possibility based structure is expanded with the help of resemblance relations assumed on the domains.

A tuple t = $\{\pi_{A1}, \pi_{A2}, \pi_{A3}\ldots, \pi_{An}\}$ of relation R ⊆ Π (D1) × Π(D2) × Π(D3) ×…× Π (Dn) with πAi □ Π (Di). The resemblance relation of domain Di is given by $resi$= (Di) × (Di) → [0, 1].

The first one is the special case of second one in terms of representation while second case deals with resemblance between possibility degrees and domain values. For fuzzy data equality based on resemblance relation following criteria must be fulfilled as in [3]: i) it should generalize resemblance relation, i.e., when two crisp values are taken as arguments, fuzzy equality and resemblance relation must

give the same result, ii) it should be reflexive, iii) it should be symmetrical.

A. Methods based on case 1:

a. Raju and Majumdar's approach: This approach does not measure degree of the equality between the two imprecise values but gives the extent to which the two fuzzy data are close to each other globally or as a whole [4]. In this approach fuzzy equality is denoted by EQ and is given as:

$$\mu_{EQ}(\pi_{A(x)}, \pi_{A(y)}) = \min_{u \in D} \Psi(\pi_{A(x)}(u), \pi_{A(y)}(u))$$

Where $\pi_{A(x)}$, $\pi_{A(y)}$ are possibility distribution for attribute A restricting its value to x and y respectively. D is the domain of attribute A and Ψ is the resemblance relation which is reflexive and symmetric over [0, 1].

b. Kerre's Approach: This approach is an extension of classical equality. It makes use of function which maps the two possibility distribution given by $\pi_{A(x)}$ and $\pi_{A(y)}$ to the possibility whether they are equal or not [5]. This method is defined as:

$$E(\pi_{A(x)}, \pi_{A(y)})(T) = sup_{u \in D} min(\pi_{A(x)}(u), \pi_{A(y)}(u))$$

$$E(\pi_{A(x)}, \pi_{A(y)})(F) = sup_{v,w \in D \text{ and } v \neq w} min(\pi_{A(x)}(u), \pi_{A(y)}(u))$$

The function E is defined as E: ∏ (D) × ∏ (D) →∏ ({T, F})

T denotes true and F shows False. The first equation shows the extent that a value exists for both A(x) and A(y). The second function shows the extent the values u for A(x) and v for A(y) belonging to D, exist such that u is dissimilar to v. If the second function results in 1 then it means that $\pi_{A(x)}$ is not equal to $\pi_{A(y)}$.

B. Methods Based on Case 2:

a. Rundensteiner's approach*:* This approach considers two possibility distributions $\pi_{A(x)}$ and $\pi_{A(y)}$ redundant or equal to each other [6], if the following two conditions hold:

a) $min_{(u,v) \in supp(\pi_{A(x)}) \times supp(\pi_{A(y)})} RES(u,v) \geq$

b) $min_{w \in D}(1 - |\pi_{A(x)}(w) - \pi_{A(y)}(w)|) \geq \beta$

Here, α □ [0,1] and β □ [0,1], supp($\pi_{A(x)}$) gives the support of $\pi_{A(x)}$. The second condition checks the resemblance between $\pi_{A(x)}$ and $\pi_{A(y)}$ when there exists a element 'd' belonging to domain 'D' such that it belongs to the support of $\pi_{A(x)}$ but not in support of $\pi_{A(y)}$ and $\pi_{A(x)}(d)$=1. For example, if $\pi_{A(x)}$(or $\pi_{A(y)}$) represents a precise value 'u' (or 'v'), the resemblance degree will be zero as soon as u ≠ v, even if 'u' and 'v' are very close in RES.

b. Chen's Approach: Kerre's [5] work is extended in this approach. Chen [7] proposed a function E_c: ∏ (D) ×∏ (D) →∏ (T, F), here T and F denote true and false respectively. The consequent possibility distribution is given as:

$$E_c(\pi_{A(x)}, \pi_{A(y)})(T) = sup_{(u,v) \in D \times D \text{ and } RES(u,v) \geq} \min(\pi_{A(x)}(u), \pi_{A(y)}(v))$$

$$E_c(\pi_{A(x)}, \pi_{A(y)})(F) = sup_{(u,v) \in D \times D \text{ and } RES(u,v) <} \min(\pi_{A(x)}(u), \pi_{A(y)}(v))$$

Where 'α' is a threshold having values in the interval [0,1]. This approach is defined to assess the possibility and the impossibility that $\pi_{A(x)} = \pi_{A(y)}$. The key idea is to perform the operations not only on identical elements but to extend the operations to close elements as well.

c. Cubero et. al's Approach: The concept of possibility and necessity defined in [8] where used to obtain the result of a query involving condition between two attributes of same tuple t. These measures are represented by Cubero in terms of strong and weak resemblance [9]. These notions can be represented as:

$$\prod(A(x) \approx_{RES} A(y)) = sup_{(v,w) \in D \times D} \min(RES(v,w), \pi_{A(x)}(v), \pi_{A(y)}(w))$$

and

$$N(A(x) \approx_{RES} A(y)) = inf_{(v,w) \in D \times D} \max(RES(v,w), 1 - \pi_{A(x)}(v), 1 - \pi_{A(y)}(w))$$

The first expression gives the weak resemblance and second expression gives the strong resemblance. Weak resemblant is the extent to which a crisp member in imprecise or vague values A(x) is resemblant to a crisp member in the imprecise value A(y). Strong Resemblance is the extent to which all crisp elements in A(x) are resemblant to all crisp elements in A(y).

d. Bosc's Approach: This approach gives the function to estimate the interchangeability that that the value A(x) can be replaced by A(y), [3]. To calculate the extent to which there exists a instance <v, $\pi_{A(y)}(v)$> in A(y) which can be substituted for u, u □ D and $\pi_{A(x)}(u) > 0$, following function is defined:

$$\mu_{s\ (A(x),A(y))}(u) = sup_{v \in supp(A(y))} \min\left(\mu_{RES}(u,v), \mu_{\theta}\left(\pi_{A(x)}(u), \pi_{A(y)}(v)\right)\right)$$

Where, $\mu_{\theta}(\alpha, \beta) = 1 - |\alpha - \beta|$.

Finally A(x) can be replaced by A(y) if all values in the support of A(x) have a substitute for A(y).

$$\mu_{repl}(A(x), A(y)) = inf_{u \in supp(A(x))} \max\left(1 - \pi_{A(x)}(u), \mu_{s(A(x),A(y))}(u)\right).$$

Since, the above function is not symmetrical, therefore

$$\mu_{EQ}(A(x), A(y)) = \min\left(\mu_{repl}(A(x), A(y)), \mu_{repl}(A(y), A(x))\right)$$

e. Z.M. Ma's Approach: The concept of semantic inclusion degree SID (π_A, π_B) and semantic equivalence degree, SED (π_A, π_B) are introduced in this approach for semantic measure of two fuzzy data π_A *and* π_B [10]. The meaning of SID (π_A, π_B) is the percentage of the semantic space of π_B which is wholly included in the semantic space of π_A, and is given by following expression:

$$SID(\pi_A, \pi_B) = \sum_{i=1}^{n} min_{u_i, u_j \in U \text{ and } Res_U(u_i, u_j) \geq} (\pi_B(u_i), \pi_A(u_i)) \Big/ \sum_{i=1}^{n} \pi_B(u_i)$$

And

SED (π_B, π_A) denote the degree that π_B *and* π_A are equivalent to each other.

$$SED(\pi_A, \pi_B) = \min (SID(\pi_A, \pi_B), SID(\pi_B, \pi_A))$$

Table 1: Comparison of the methods

Approaches	Key Idea	Equivalence function	Nature of function	Interval	Type	Whether based on resemblance relation or not.
a) Raju & Majumdar's approach[4]	To measure the extent to which two representations are close to each other.	μ_{Eq}	Reflexive and Symmetric	[0,1]	Based on identity i.e. degree of each element in the domain is matched.	Yes, here resemblance gives the strict equality between the elements of domain.
b) Kerre's Approach[5]	Gives the extent to which a value possible for both A(x) and A(y).	E: ∏(D) × ∏(D) → ∏({T,F})	-	[0,1]	Based on identity.	Same as above.
c) Rundenst--einer's Approach[6]	Gives the measure that two fuzzy data are considered α-β redundant or not.	Inequality conditions as defined above should hold.	-	[0,1] for both α and β.	Based on resemblance.	Yes, here resemblance relation gives the fuzzy equality between the values in Domain.
d) Chen's Approach[7]	Assess the possibility and impossibility that $\pi_{A=} \pi_B$. The idea to perform operations not only on identical elements but also on close elements.	E_c: ∏(D) × ∏(D) →∏(T, F)	Does not generalize the resemblance relation.	[0,1]	Based on resemblance.	Same as above.
e) Cubero et. al's Approach[8]	Based on strong and weak resemblance notions for the two fuzzy values being approx. equal.	$\prod(A(x) \approx_{RES} A(y))$ gives weak resemblance and $N(A(x) \approx_{RES} A(y))$ gives strong resemblance.	Strong resemblance is not reflexive.	[0,1]	Based on resemblance.	Same as above.
f) Bosc's Approach[3]	Measures the interchangeability of two fuzzy values.	μ_{Eq}	Reflexive and symmetrical.	[0,1]	Based on resemblance.	Yes, and also on another measure called proximity measure θ.
g) Z.M. Ma's Approach[10]	Gives the semantic measure of the two fuzzy values.	$SED(\pi_A, \pi_B)$	Symmetrical.	[0,1]	Based on semantic space of fuzzy data.	Yes.

III. COMPARISON OF METHODS

The comparison of methods has been done in table 1. The restriction on approaches based on identity relation is that strict equality is used to compare the values whereas in the approaches based on resemblance have resemblance relation based on fuzzy equality between the values of domain. The approach (c) suffers for the counter intuitive problem i.e. the fuzzy data with same representation can said not equal as the two criteria have been independently set. The approach (f) is an extension of (c) but the counter intuitive problem still exists in (f). In (d) some inconsistencies are present and do not give the extent to which the fuzzy values are interchangeable. In approach (e) weak resemblance is too optimistic and strong resemblance is too severe for fuzzy data assessment, (d) is similar to weak resemblance except that the crisp values are used to calibrate the set of comparable values.

IV. CONCLUSION

The study of measures for comparing fuzzy data has been done. The various approaches have been compared and some problems in these approaches have been listed. These approaches can be used for finding the redundancy in the fuzzy databases. These approaches can be used for defining the fuzzy functional dependencies which can be implemented in the form of constraints in fuzzy databases. These can be useful in integration of fuzzy relations for entity identification in instance integration process. Finally, these methods can be used in extending fuzzy querying by introducing new operators for fuzzy data comparison.

V. REFERENCES

[1] P. Lim, J. Srivastava, S. Prabhakar, J. Richardson, "Entity identification problem in database integration", Proc, Intl. Conf. on Data Enggineering, 1993, pp. 154-163.

[2] A.K. Sharma, A. Goswami, D.K. Gupta, "Integration of fuzzy databases: Problems & Solutions", International Journal of Computer Applications (0975-8887), Vol. 2- No. 3, 2010.

[3] P. Bosc and O. Pivert, "On the comparison of imprecise values in fuzzy databases", Proceedings of the 6th IEEE International Conference on fuzzy systems, Vol. 140, 2003, pp.207-227.

[4] K.V.S.V.N. Raju, A.K. Majumdar, "Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems". ACM Transactions on Database Systems, 13(2), 1988, 129-166.

[5] E.E. Kerre, "Fuzzy sets and approximate reasoning", Lecture notes for the course (Special Topics in Computer Sciences). Lincoln, NB: University of Nebraska, 1988.

[6] E.A. Rundensteiner, L.W. Hawkes, W. Bandler, "Onnearness measures in fuzzy relational data models", International Journal of Approximate Reasoning, 3, 1989, pp. 267-98.

[7] G.Q. Chen, E.E. Kerre, J. Vandenbulcke, "A general treatment of data redundancy in a fuzzy relational data model", Journal of the American Society for Information Science. Vol. 43, 1992, pp. 304-311.

[8] H.Prade, C.Testemale, "Generalizing database relational algebra for the treatmrnt of incomplete or uncertain information and vague queries". Information Sciences, 34, 1984, 115-143.

[9] J.C. Cubero, O. Pons, M.A. Vila, "Weak and strong resemblance in fuzzy functional dependencies", Proc FUZZ IEEE' 94, 1994, 162-166.

[10] Z.M. Ma, W. J. Zhang, and W. Y. Ma, "Semantic measure of fuzzy data in extended possibility-based fuzzy relational databases", International Journal of Intelligent Systems, Vol. 15, 2000, pp. 705-716.

