Pay-As-You-Go Data Integration Using Functional Dependencies

Naser Ayat[#1], Hamideh Afsarmanesh[#2], Reza Akbarinia[*3], Patrick Valduriez[*4]

[#]Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
[1]s.n.ayat@uva.nl, [2]h.afsarmanesh@uva.nl
[*]INRIA and LIRMM, Montpellier, France
[3,4]{Firstname.Lastname@inria.fr}

Abstract. Setting up a full data integration system for many application contexts, e.g. web and scientific data management, requires significant human effort which prevents it from being really scalable. In this paper, we propose IFD (Integration based on Functional Dependencies), a pay-as-you-go data integration system that allows integrating a given set of data sources, as well as incrementally integrating additional sources. IFD takes advantage of the background knowledge implied within functional dependencies for matching the source schemas. Our system is built on a probabilistic data model that allows capturing the uncertainty in data integration systems. Our performance evaluation results show significant performance gains of our approach in terms of recall and precision compared to the baseline approaches. They confirm the importance of functional dependencies and also the contribution of using a probabilistic data model in improving the quality of schema matching. The analytical study and experiments show that IFD scales well.

Keywords: data integration, uncertain data integration, functional dependency

1 Introduction

Data integration systems offer uniform access to a set of autonomous and heterogeneous data sources. Sources may range from database tables to web sites, and their numbers can range from tens to thousands. The main building blocks of a typical data integration application are mediated schema definition, schema matching and schema mapping. The mediated schema is the schema on which users pose queries. Schema matching is the process of finding associations between the elements (often attributes or relations) of different schemas, e.g. a source schema and the mediated schema in the popular Local As View (LAV) approach [1]. Schema mapping (also referred to as semantic mapping) is the process of relating the attributes of source schemas to the mediated schema (sometimes using expressions in a mapping language). The output of schema matching is used as input to schema mapping algorithms [1].

Setting up a full data integration system with a manually designed mediated schema requires significant human effort (e.g. domain experts and database

designers). On the other hand, there are many application contexts, e.g. web, scientific data management, and personal information management, which do not require full integration to provide useful services [2]. These applications need to start with a data integration application in a complete automatic setting for reducing human effort and development time and put more effort on improving it as needed. Let us present a motivating example from the scientific data management context.

Example 1. Consider a researcher who is interested in the less-known or yet unknown functions of the protein ABCC8 related to diabetes. While biological experiments are the ultimate means for verifying predicted functions, she must first discover and suggest such functions. For doing this, she should perform manual exploratory searches over numerous online sources. For example, she should consider both well-known databases such as EntrezGene, EntrezProtein and less-known databases of other research labs as well. Having a data integration system with approximate answers can considerably save the time and reduce the research cost in this domain. It is sufficient to set up such a system in a complete automatic setting and spend more effort to improve it only if it is necessary. This recent setting, referred to by pay-as-you-go data integration, has attracted considerable attention, e.g. [2–5]. The ultimate goal of this setting is to reduce human burden, and thereby reduce the time and cost of data integration while providing sufficient integration [2].

The goal of our work is to provide a pay-as-you-go data integration system that deals with the uncertainty arising during the matching process. To capture the uncertainty, we generate Probabilistic Mediated Schemas (PMSs) which have shown to be promising [6]. The idea behind PMSs is to have several mediated schemas, each one with a probability that indicates the closeness of the corresponding mediated schema to the ideal mediated schema.

The closest related work to ours is that of Sarma et al. [3] which based on PMSs proposed UDI (Uncertain Data Integration), an uncertain data integration system. However, UDI may fail to capture some important attribute correlations, and thereby produce low quality answers. Let us clarify this by an example which is the same as the running example in [3].

Example 2. Consider the following schemas both describing people:
$S_1(name, hPhone, hAddr, oPhone, oAddr)$
$S_2(name, phone, address)$
In S_2, the attribute *phone* can either be a home phone number or an office phone number, and the attribute *address* can either be a home address or an office address.

An ideal data integration system should capture the correlation between $hPhone$ and $hAddr$ and also between $oPhone$ and $oAddr$. Specifically, it must generate schemas which group the *address* and $hAddr$ together if *phone* and $hPhone$ are grouped together. Similarly it should group the *address* and $oAddr$ together if *phone* and $oPhone$ are grouped together. In other words either of the following schemas should be generated (we abbreviate $hPbone, oPhone, hAddr, oAddr$ as hP, oP, hA, and oA respectively):

$M_1(\{name, name\}, \{phone, hP\}, \{oP\}, \{address, hA\}, \{oA\})$
$M_2(\{name, name\}, \{phone, oP\}, \{hP\}, \{address, oA\}, \{hA\})$

Although these schemas are generated by UDI, they are overwhelmed by schemas in which the attribute correlations are not respected. Thus, by producing a large number of schemas which can easily be exponential, the desirable schemas get a very low probability. This occurs because UDI does not consider attribute correlations. Most attribute correlations are expressed within Functional Dependencies (FDs). For example let F_1 and F_2 be the set of FDs of S_1 and S_2 respectively:

$F_1 = \{hPhone \rightarrow hAddr, oPhone \rightarrow oAddr\}$
$F_2 = \{phone \rightarrow address\}$

These FDs show the correlation between attributes. For example, $hPhone \rightarrow hAddr$ indicates that the two attributes $hPhone$ and $hAddr$ are correlated. Considering the pairs of FDs from different sources can help us extracting these correlations and achieving the goal of generating mediated schemas that represent these correlations. For example, the FD pair $phone \rightarrow address$ and $hPhone \rightarrow hAddr$ indicates that if we group $phone$ and $hPhone$ together, we should also group $address$ and $hAddr$ together, as well as $oPhone$ and $oAddr$.

In this paper, we propose IFD (Integration based on Functional Dependencies), a pay-as-you-go data integration system that takes into account attribute correlations by using functional dependencies, and captures uncertainty in mediated schemas using a probabilistic data model. We model the schema matching problem as a clustering problem with constraints. This allows us to generate mediated schemas using algorithms designed for the latter problem. In our approach, we build a custom distance function for representing the knowledge of attribute semantics which we extract from FDs. We also propose a new metric (i.e. FD-point) for ranking the generated mediated schemas in the clustering process, and selecting high quality ones. IFD allows integrating a given set of data sources, as well as incrementally integrating additional sources, without needing to restart the process from scratch. To validate our approach, we implemented IFD as well as baseline solutions. The performance evaluation results show significant performance gains of our approach in terms of recall and precision compared to the baseline approaches. They confirm the importance of FDs in improving the quality of uncertain mediated schemas.

The rest of the paper is organized as follows. In Section 2, we make our assumptions precise and define the problem. In Section 3, we propose IFD, and describe its architecture, components and algorithms. We also analyze the execution cost of IFD's algorithms. Section 4 describes our performance validation. Section 5 discusses related work, and Section 6 concludes.

2 Problem Definition

In this section, we first give our assumptions and some background about PMSs. Then, we state the problem we address in this paper.

For the applications which we consider (e.g., scientific data management), we assume the availability of functional dependencies for the attributes of sources. This is a reasonable assumption in the applications which we consider, in particular scientific applications, because the data source providers are willing to provide the full database design information, including functional dependencies. However, there are contexts such as the web in which functional dependencies are not available. For these applications, we can use one of the existing solutions, e.g. [7, 4] to derive functional dependencies from data. Another assumption, which we make for ease of presentation, is that the data model is relational.

Now, we define some basic concepts, e.g. functional dependencies and mediated schemas, and then state the problem addressed in this paper. Let S be a set of source schemas, say $S = \{S_1, \ldots, S_n\}$, where for each $S_i, i \in [1, n], S_i = \{a_{i,1}, \ldots, a_{i,l_i}\}$, such that $a_{i,1}, \ldots, a_{i,l_i}$ are the attributes of S_i. We denote the set of attributes in S_i by $att(S_i)$, and the set of all source attributes as A. That is $A = \cup_i att(S_i)$. For simplicity, we assume that S_i contains a single table. Let F be the set of functional dependencies of all source schemas, say $F = \{F_1, \ldots, F_n\}$. For each $S_i, i \in [1, n]$, let F_i be the set of functional dependencies among the attributes of S_i, i.e. $att(S_i)$, where each $fd_j, fd_j \in F_i$ is of the form $L_j \rightarrow R_j$ and $L_j \subseteq att(S_i), R_j \subseteq att(S_i)$. In every F_i, there is one fd of the form $L_p \rightarrow R_p$, where $R_p = att(S_i)$, i.e. L_p is the primary key of S_i.

We assume that every attribute in the data sources can be matched with at most one attribute in other data sources, which means we only consider one-to-one mappings. We do this for simplicity and also because this kind of mapping is more common in practice. For a set of sources S, we denote by $M = \{A_1, \ldots, A_m\}$ a mediated schema, where $A_i \subseteq A$, and for each $i, j \in [1, m], i \neq j \Rightarrow A_i \cap A_j = \emptyset$. Each attribute involved in A_i is called a mediated attribute. Every mediated attribute ideally consists of source attributes with the same semantics.

A probabilistic mediated schema (PMS) for a set S of source schemas is the set $N = \{(M_1, P(M_1)), \ldots, (M_k, P(M_k))\}$ where $M_i, i \in [1, k]$, is a mediated schema, and $P(M_i)$ is its probability. For each $i, j \in [1, k], i \neq j \Rightarrow M_i \neq M_j$,i.e. M_i and M_j are different clusterings of $att(S)$; and $\sum_{i=1}^{k} P(M_i) \leq 1$.

Since each mediated schema corresponds to a clustering of source attributes, we can measure its quality by computing the F-measure of the clustering.

Let us now state the problem we address. Suppose we are given a set of source schemas S, and a set of functional dependencies F and a positive integer number k as input. Our problem is to efficiently find a set of k probabilistic mediated schemas which have the highest F-measure.

3 Data Integration Based on Functional Dependencies

In this section, we describe IFD, a data integration system that automatically performs the tasks of mediated schema generation and the attribute matching, by taking advantage of functional dependencies among the source attributes. In the rest of this section, we first briefly describe the architecture of our data integration system. Then, we describe our approach for schema matching.

3.1 System Architecture

Figure 1 depicts the architecture of our system, which consists of two main parts of schema matching and query processing, in part A and part B respectively. The components of part A operate during the set-up time of the system and the components of part B operate at query evaluation time. In this paper, our focus is on the schema matching part (part A) but we include components of part B in the architecture of our system to provide a complete picture of a data integration system. A more detailed description of the components is available in the extended version of this paper [8].

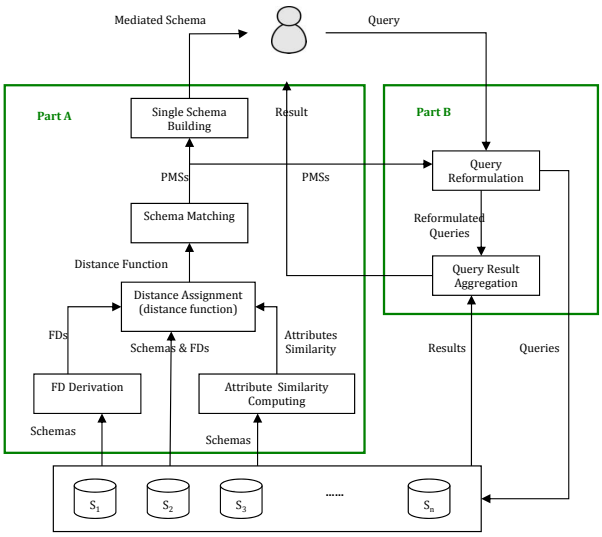

Fig. 1. Architecture of our data integration system

To build the mediated schema automatically, we cluster the source attributes by putting semantically equivalent attributes in the same cluster. We use a clustering algorithm that works based on a *distance matrix* (i.e. the distance between every two attributes). Specifically we use the single-link CAHC (Constrained Agglomerative Hierarchical Clustering) algorithm [9]. To assign the distances between the attributes, we use the attributes' name similarity as well as some heuristics we introduce about FDs.

3.2 FD Heuristics

We use heuristic rules related to FDs in order to assign the distance of attributes. Before describing our heuristics, let us first define *Match* and *Unmatch* concepts. Consider a_1 and a_2 as two typical attributes. If we want to increase their chance of being put in the same cluster, we set their distance to *MD* (i.e. Match Distance) which is 0 or a number very close to 0. In this case, we say that we matched a_1 with a_2, and we show this by $Match(a_1, a_2)$. In contrast, if

we want to decrease their chance of being put in the same cluster, then we set their distance to *UMD* (i.e. Un-Match Distance) which is 1 or a number very close to 1. In this case, we say that we *unmatched* a_1 and a_2 and we show this by $Unmatch(a_1, a_2)$. Now, Let us use the following example to illustrate the heuristics.

Example 3. Consider two source schemas, both describing a university course schedule. In this example, primary keys are underlined; F_1 and F_2 are the sets of FDs of S_1 and S_2 respectively:

$S_1(\underline{term, c\#, section\#}, coursename, instructor, name, time, room)$
$S_2(\underline{semester, course, sec\#}, name, instructor, ins\_name, location)$
$F_1 = \{c\# \rightarrow coursename, instructor \rightarrow name\}$
$F_2 = \{course \rightarrow name, instructor \rightarrow ins\_name\}$

Heuristic 1 *Let S_p and $S_q, p \neq q$, be two source schemas. Then,*

$$Match(a_{p,i}, a_{q,k}) \Rightarrow unmatch(a_{p,i}, a_{q,l}) \wedge unmatch(a_{q,k}, a_{p,j})$$

where $a_{p,i} \in att(S_p), a_{p,j} \in att(S_p) \setminus \{a_{p,i}\}, a_{q,k} \in att(S_q), a_{q,l} \in att(S_q) \setminus \{a_{q,k}\}$.

The reason behind heuristic 1 is that each attribute can be matched with at most one attribute of the other source.

Heuristic 2 *Let $fd_p : a_{p,i} \rightarrow a_{p,j}$ and $fd_q : a_{q,k} \rightarrow a_{q,l}$ be two FDs, where $fd_p \in F_p, fd_q \in F_q, p \neq q$. Then, $similarity(a_{p,i}, a_{q,k}) > t_L \Rightarrow Match(a_{p,j}, a_{q,l})$ where t_L is a certain threshold and similarity is a given similarity function.*

The reason behind heuristic 2 is that we consider the set of facts that the two sources are assumed to be from the same domain, and both attributes $a_{p,j}$ and $a_{q,l}$ are functionally determined by the attributes $a_{p,i}$, and $a_{q,k}$ respectively, which themselves have close name similarity. Thus, we heuristically agree that: the probability of $Match(a_{p,j}, a_{q,l})$ is higher than that of $Match(a_{p,j}, a_{q,s})$ and $Match(a_{q,l}, a_{p,r})$, where $a_{q,s} \in att(S_q) \setminus \{a_{q,l}\}$ and $a_{p,r} \in S_p \setminus \{a_{p,j}\}$. Therefore, in such a case we match $a_{p,j}$ with $a_{q,l}$ to reflect this fact. Note that this heuristic has a general form in which there are more than one attribute on the sides of the FDs (see Section 3.3).

By applying heuristic 2 on Example 3, we have the FD $instructor \rightarrow name$ from S_1, and $instructor \rightarrow ins\_name$ from S_2. There is only one attribute at the left side of these FDs, and their name similarity is equal to 1 that is the maximum similarity value. Thus, we match the *name* with the *ins_name* which appear on the right side of these FDs. Notice that in this example, FDs guided us to recognize that the *name* in S_2 is in fact the instructor's name, and not the course's name. This kind of mistake is typically made by approaches which only rely on name similarity for attribute matching.

Heuristic 3 *Let PK_p and $PK_q, p \neq q$, be the primary keys of S_p and S_q respectively. Then,*

$$(\exists a_{p,i} \in PK_p, a_{q,j} \in PK_q \mid (a_{p,i}, a_{q,j}) = \underset{a_p \in PK_p, a_q \in PK_q}{\arg\max} similarity(a_p, a_q)) \wedge$$
$$(similarity(a_{p,i}, a_{q,j}) > t_{PK}) \Rightarrow Match(a_{p,i}, a_{q,j})$$

where t_{PK} is a certain threshold and similarity is a given similarity function.

The reason behind heuristic 3 is simple. Since we assume sources are from the same domain, there are a number of specific attributes which can be part of the primary key. Although these attributes may have different names in different sources, it is reasonable to expect that some of these attributes from different sources can be matched together. Obviously, we can set t_{PK} to a value less than the value we set for t_L because typically the probability of finding matching attributes in the primary key attributes is higher than the other attributes. After matching $a_{p,i}$ with $a_{q,j}$, we remove them from PK_p and PK_q respectively, and continue this process until the similarity of the pair with the maximum similarity is less than the threshold t_{PK} or one of the PK_p or PK_q has no more attributes to match.

Now we apply heuristic 3 to Example 3. It is reasonable to match the attributes: *term*, *c#*, and *section#* of S_1 with *semester*, *course*, and *sec#* of S_2 rather than with other attributes of S_2, and vice versa. The attribute pair with the maximum similarity is (*section#*, *sec#*). If we choose a good threshold, we can match these attributes together. The similarity of other attribute pairs is not high enough to pass the wisely selected threshold values.

Heuristic 4 *Let PK_p and $PK_q, p \neq q$, be the primary keys of S_p and S_q respectively. Then,*

$$(\exists a_{p,i} \in PK_p, a_{q,j} \in PK_q, fd_p \in F_p, fd_q \in F_q \mid fd_p : a_{p,i} \rightarrow R_p, fd_q : a_{q,j} \rightarrow R_q) \Rightarrow Match(a_{p,i}, a_{q,j}) \quad (1)$$

and also

$$(RHS(1) \wedge R_p = \{a_{p,r}\} \wedge R_q = \{a_{q,s}\}) \Rightarrow Match(a_{p,r}, a_{q,s}) \quad (2)$$

We can apply heuristic 4 when we have two attributes in two primary keys which each of them is the single attribute appearing at the left side of a FD. In this case, we match these attributes with each other(rule 1). We also match the attributes on the right sides of the two FDs if there is only one attribute appearing at the right side of them (rule 2).

By applying heuristic 4 on Example 3, we match *c#* with *course* which is a right decision. We do this because of the two FDs: *c#* $\rightarrow$ *coursename* and *course* $\rightarrow$ *name*. We also match *coursename* with *name* which are the only attributes appearing at the right side of these FDs. Had we used name similarity only, we would have very likely matched *coursename* with *course* for example, which is a wrong decision.

Heuristic 5 *Let PK_p and $PK_q, p \neq q$, be the primary keys of S_p and S_q respectively. Then,*

$$(\forall a_{p,r} \in PK_p \setminus \{a_{p,i}\}, \exists a_{q,s} \in PK_q \setminus \{a_{q,j}\} \mid Match(a_{p,r}, a_{q,s})) \wedge (|PK_p| = |PK_q|) \Rightarrow Math(a_{p,i}, a_{q,j})$$

Algorithm 1 Distance Assignment

Input: 1) Source schemas $S_1, \ldots, S_n$; 2) The sets of FDs $F_1, \ldots, F_n$ (the FDs related to PK are omitted); 3) $P = \{PK_1, \ldots, PK_n\}$ The set of primary keys of all sources.
Output: Distance matrix $D[m][m]$.

1: compute $A = \{a_1, \ldots, a_m\}$ the set of all source attributes
// match attributes on the right sides of FDs
2: **for all** FD pair $fd_i \in F_k, fd_j \in F_l, k \neq l$ **do**
3: **if** $IsMatch(L_i, L_j)$ **then**
4: make local copies of fd_i, fd_j
5: find the attribute pair $a_p \in R_i, a_q \in R_j$ with the maximum similarity s
6: **if** $s > t_R$ **then**
7: $DoMatch(a_p, a_q)$
8: $R_i \leftarrow R_i \setminus \{a_p\}; R_j \leftarrow R_j \setminus \{a_q\}$
9: **if** $\mid R_i \mid > 0$ and $\mid R_j \mid > 0$ **then**
10: go to 5
// match PK attributes
11: **for all** pair $PK_i, PK_j \in P$, where they are PKs of S_i and S_j respectively **do**
12: make local copies of PK_i and PK_j
13: **for all** pair $a_p \in PK_i, aq \in PK_j$ **do**
14: **if** $\exists fd_k \in F_i$ and $fd_l \in F_j$ such that $L_k = \{a_p\}$ and $L_l = \{a_q\}$ **then**
15: $DoMatch(a_p, a_q)$
16: $PK_i \leftarrow PK_i \setminus \{a_p\}$; $PK_j \leftarrow PK_j \setminus \{a_q\}$
17: **if** $R_k = \{a_s\}$ and $R_l = \{a_t\}$ **then**
18: $DoMatch(a_p, a_q)$
19: find the attribute pair $a_p \in PK_i$ and $a_q \in PK_j$ with maximum similarity s
20: **if** $s > t_{PK}$ **then**
21: $DoMatch(a_p, a_q)$
22: $PK_i = PK_i \setminus \{a_p\}$; $PK_j = PK_j \setminus \{a_q\}$
23: **if** $\mid PK_i \mid > 0$ and $\mid PK_j \mid > 0$ **then**
24: go to 19
25: **if** $PK_i = \{a_p\}$ and $PK_j = \{a_q\}$ **then**
26: $DoMatch(a_p, a_q)$
27: **for all** attribute pair $a_i, a_j \in A$ which $D[a_i][a_j]$ has not been computed yet **do**
28: **if** $a_i, a_j \in S_k$ (the same source) **then**
29: $D[a_i][a_j] \leftarrow UMD$
30: **else**
31: $D[a_i][a_j] \leftarrow similarity(a_i, a_j)$
32: $\forall a_i, a_j, a_k \in A$ **if** $(D[a_i][a_k] = MD$ and $D[a_k][a_j] = UMD)$ **then** $D[a_i][a_j] \leftarrow UMD$
33: $\forall a_i, a_j, a_k \in A$ **if** $(D[a_i][a_k] = MD$ and $D[a_k][a_j] = MD)$ **then** $D[a_i][a_j] \leftarrow MD$
34: $\forall a_i, a_j \in AD[a_i][a_j] \leftarrow D[a_j][a_i]$

We can apply heuristic 5 when all attributes of PK_p and PK_q have been matched, and only one attribute is left in each of them. We match these two attributes with each other hoping that they are semantically the same. Coming back to Example 3, there is only one attribute left in each of the primary keys that we have not yet matched (i.e. *term*, *semester*) that we can match using this heuristic.

3.3 Distance Assignment Algorithm

Algorithm 1 describes how we assign distances to attribute pairs and build the distance matrix that is used in schema matching. Steps 2-10 of the algorithm find FD pairs from different sources which their left sides match together and then try to match attribute pairs on the right sides of these FDs. Steps 5-7 find the attribute pairs (a_p, a_q) whose similarity is maximum. If the similarity of a_p and a_q is more than threshold t_R, their distance is set to MD (Match Distance), and the distances between each of them and any other source-mates are set to UMD

(Unmatch Distance). The algorithm uses the $DoMatch$ procedure for matching and unmatching attributes. It gets the attributes which should be matched as parameter, matches them, and unmatches every one of them with the other ones' source-mates. Generally, whenever the algorithm matches two attributes with each other, it also unmatches the two of them with the other one's source-mates because every attribute of a source can be matched with at most one attribute of every other source. Steps 8-10 remove the matched attributes from the list of unmatched attributes, and repeat the matching process if there are still some attributes remaining for matching.

Step 3 uses the $IsMatch$ function. This function takes as parameter the left sides of two FDs and returns true if they can be matched together, otherwise it returns false. It first checks whether the input parameters are two sets of the same size. Then, it finds the attribute pair with maximum name similarity and treats it as matched pair by removing the attributes from the list of unmatched attributes if their similarity is more than threshold t_L. It repeats the matching process until there is no more attribute eligible for matching. After the matching loop is over, the function returns true if all attribute pairs have been matched together, otherwise it returns false which means the matching process has not been successful.

Notice that we do not reflect the matching of attributes of the left sides of FDs in the distance matrix. The reason is that for these attributes (in contrast to those on the right side), the matching is done just based on attribute name similarity and not the knowledge in FDs.

In this algorithm, we use three different similarity thresholds (i.e. t_L, t_R, and t_{PK}). We do this to have more flexibility in the matching. The discussion on setting these parameters is available in the extended version of this paper[8].

Coming back to Algorithm 1, steps 11-26 apply PK heuristics to every PK pair and try to match their attributes based on these heuristics. Steps 13-18 check every attribute pair of two PKs to see if they are the only attributes at the left sides of two FDs. If yes, then these attributes are matched together. Steps 19-24 find the attribute pair with the maximum name similarity and if it is more than threshold t_{PK}, the attributes are matched together. The matching process continues until there is at least one attribute in every PK and the similarity of the attribute pair with the maximum similarity is more than threshold t_{PK}. After the matching process, if each of the two PKs has only one attribute left, their attributes are matched with each other by steps 25-26.

Steps 27-31 set the distances of attribute pairs which have not been computed by the heuristic rules. Step 28 checks if the attributes are from the same source, in which case their distance is set to UMD; otherwise the distance is set to their name similarity by step 31.

Steps 32-33 perform a transitive closure over the match and unmatch constraints. Step 34 deals with the symmetric property of the distance function to ensure that the returned distance is independent from the order of attributes.

Algorithm 2 Schema Matching

Input: 1) Source schemas $S_1, \ldots, S_n$; 2) Distance matrix $D[m][m]$; 3) Number of needed mediated schemas k.
Output: A set of probabilistic mediated schemas.

1: compute $A = \{a_1, \ldots, a_m\}$ the set of all source attributes
2: let C be the set of clusters c_i such that $c_i = \{a_i\}, a_i \in A, i \in [1, m]$
3: $M \leftarrow C$
4: find two clusters $c_i, c_j \in C$ having the minimum distance d_{min} while distance d_{ij} between c_i and c_j is computed as follows:
5: **if** $\exists a_k \in c_i, a_l \in c_j, a_k, a_l \in S_p$ **then**
6: $d_{ij} \leftarrow \infty$
7: **else**
8: $d_{ij} \leftarrow Min(D[a_k][a_l]), a_k \in c_i, a_l \in c_j$
9: **if** $d_{min} \neq \infty$ **then**
10: merge c_i with c_j
11: Add the newly added mediated schema to M
12: go to 4
13: **for each** $C_i \in M$ compute the $FDpoint_i$ as the number of attribute pairs recommended by distance matrix and respected by C_i
14: $FDpoint_{max} \leftarrow Max(FDpoint_i), C_i \in M$
15: $M \leftarrow \{C_i \mid C_i \in M, FDpoint_i = FDpoint_{max}\}$
16: **if** $k < |M|$ **then**
17: select k mediated schemas randomly from M
18: assign probability $\frac{1}{k}$ to every selected mediated schema and return them
19: **else**
20: assign probability $\frac{1}{|M|}$ to every $C_i \in M$ and return them

3.4 Schema Matching Algorithm

The distances between attributes are used for computing the distance between clusters in the clustering method, i.e. CAHC. Algorithm 2 describes how we create probabilistic mediated schemas. This algorithm takes as input the source schemas, distance matrix, and the needed number of mediated schemas (k) which is specified by the user. Steps 1-2 create the first mediated schema by putting every attribute in a cluster. The algorithm stores all created mediated schemas in the set M, and so does for the first created mediated schema in step 3.

Steps 4-8 look for the two clusters with the minimum distance while the distance between two clusters is defined as follows: if the clusters have two attributes from the same source, the distance between them is infinity; otherwise the minimum distance between two attributes, each from one of the two clusters, is regarded as the distance between the two clusters. Steps 9-12 merge these clusters together and store this newly created mediated schema in M and continues this process by going to step 4. The necessary condition for merging clusters is that their distance should not be equal to infinity. We get the infinity as the minimum distance between clusters when every two clusters have attributes from the same source. In such a case, we stop creating the mediated schemas.

Since for all generated mediated schemas we do not let unmatched attributes to be put in the same cluster, we count the number of matched pairs which has been respected by the mediated schema, as a metric for ranking mediated schemas. We call this metric the FD-point. For every created mediated schema, Step 13 computes its FD-point, which is a metric for measuring the quality of mediated schemas and for selecting only the high quality ones. Distance matrix recommends some attribute pairs to be put in the same cluster by returning

their distance as MD. FD-point is defined as the number of these recommendations which are respected by the mediated schema. Steps 14-15 select the mediated schemas with the maximum FD-point. We call them as eligible mediated schemas.

Steps 16-20 return k randomly selected mediated schemas to the user. Since the algorithm has no means for differentiating between eligible mediated schemas, it assigns equal probabilities to all returned mediated schemas.

Let m be the number of the attributes of all sources, then the running time of algorithms 1 and 2 together is $\theta(m^3)$. The details about the complexity analysis of our algorithms are available in the extended version of this paper [8]).

IFD starts with a given set of sources and ends up generating several PMSs from these sources. A useful property of IFD is that it allows new sources to be added to the system on the fly. The details of this process are available in the extended version of this paper [8].

4 Performance Evaluation

In this section, we study the effectiveness of our data integration solution. In particular, we show the effect of using functional dependencies on the quality of generated mediated schemas. We compare our solution with the one presented in [3] which is the closest to ours. To examine the contribution of using a probabilistic approach, we compare our approach with two traditional baseline solutions that do not use probabilistic techniques, i.e. they generate only one single deterministic mediated schema.

The rest of this section is organized as follows. We first describe our experimental setup. Then we compare the performance of our solution with the competing approaches.

4.1 Experimental Setup

We implemented our system (IFD) in Java. We took advantage of Weka 3-7-3 classes [10] for implementing the hierarchical clustering component. We used the SecondString tool[1] to compute the Jaro Winkler similarity [11] of attribute names in pair-wise attribute comparison. We conducted our experiments on a Windows XP machine with Intel core 2 GHz CPU and 2GB memory.

In our experiments, we set the number of mediated schemas (denoted as n) to 1000, which is relatively high, in order to return all eligible mediated schemas. Our experiments showed similar results when we varied n considerably (e.g. n = 5). The default values for the parameters of our solution are as follows. We set similarity threshold for PK attributes (t_{PK}) to 0.7, similarity threshold for attributes on the left side of functional dependencies (t_L) to 0.9, similarity threshold for attributes on the right side of functional dependencies (t_R) to 0.8, the distance between attributes being matched (MD) to 0, and the distance between attributes being unmatched (UMD) to 1.

[1] Secondstring. `http://secondstring.sourceforge.net/`

We evaluated our system using a dataset in the university domain. This dataset[2] consists of 17 single-table schemas which we designed ourselves. For having variety in attribute names, we used Google Search with "computer science" and "course schedule" keywords and picked up the first 17 related results. For every selected webpage, we designed a single-table schema which could be the data source of the course schedule information on that webpage and we used data labels as attribute names of the schema. Also, we created primary key and functional dependencies for every schema using our knowledge of the domain.

To evaluate the quality of generated mediated schemas, we tested them against the mediated schema which we created manually. Since each mediated schema corresponds to a clustering of source attributes, we measured its quality by computing the precision, recall, and F-measure of the clustering. We computed the metrics for each individual mediated schema, and summed the results weighted by their respective probabilities.

To the best of our knowledge, the most competing approach to ours (IFD) is that of Sarma et al. [3] which we denote by UDI as they did. Thus, we compare our solution with UDI as the most competing probabilistic approach. We implemented UDI in Java. We used the same tool in our approach for computing pair-wise attribute similarity as in UDI. Also, we set the parameters edge-weight threshold and error bar to 0.85 and 0.02 respectively. Since the time complexity of UDI approach is exponential to the number of uncertain edges, we selected the above values carefully to let it run.

To examine the performance gain of using a probabilistic technique, we considered two baseline approaches that create a single mediated schema:

- FD1: creates a deterministic mediated schema as follows. In Algorithm 2, we count the number of FD recommendations and obtain the maximum possible FD-point, then we stop at the first schema which gets this maximum point.
- SingleMed: creates a deterministic mediated schema based on Algorithm 4.1 in [3]. We set frequency threshold to 0 and the edge weight threshold to 0.85.

Also, to evaluate the contribution of using functional dependencies in the quality of generated mediated schemas, we considered Algorithm 2 without taking advantage of the FD recommendations (WFD) and compared it to our approach.

4.2 Results

Quality of Mediated Schemas In this section, we compare the quality of mediated schemas generated by our approach (IFD) with the ones generated by UDI and other competing approaches.

Figure 2 compares the results measuring precision, recall, and F-measure of IFD, UDI, Single-Med, FD1, and WFD. It shows that IFD obtains better results than UDI. It improves precision by 23%, recall by 22%, and F-measure by 23%.

[2] The dataset is available at `http://www.science.uva.nl/CO-IM/papers/IFD/IFD-test-dataset.zip`

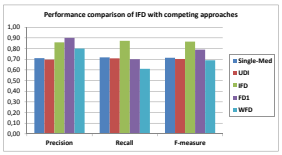

Fig. 2. Performance comparison of IFD with competing approaches

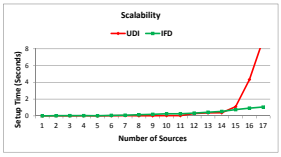

Fig. 3. Execution time comparison of IFD and UDI (seconds)

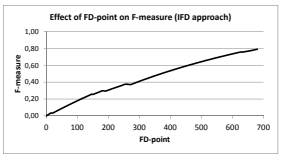

Fig. 4. Effect of FD-point on F-measure in IFD approach

Figure 2 also shows the contribution of using FD recommendations in the quality of the results. WFD (Without FD) shows the results of our approach without using FD recommendations. It is obvious that using these recommendations has considerable effect on the results.

Furthermore, Figure 2 shows the performance gain of using a probabilistic approach rather than a single deterministic schema approach. FD1 applies all of the FD recommendations to obtain the mediated schema with the maximum FD-point, then stops and returns the resulted mediated schema. On the other hand, IFD does not stop after applying all FD recommendations but since there is no further FD recommendation, it starts merging clusters based on the similarity of their attribute pairs. This increases recall considerably, but reduces precision a little because some pairs are clustered wrongly. Overall, IFD improves F-measure by 8% compared to FD1. On the other hand, this Figure shows that UDI does not get such performance gain compared to Single-Med which creates a single deterministic schema. This happens because UDI cannot select the high quality schemas among the generated schemas.

Scalability To investigate the scalability of our approach, we measure the effect of the number of sources (n) on its execution time. By execution time, we mean the setup time needed to integrate n data sources. For IFD, the execution time equals to the execution time of computing distances using Algorithm 1 plus the execution time of generating mediated schemas using Algorithm 2. For UDI, we only consider the time needed to generate mediated schemas to be fair in our comparison. For UDI, the execution time is the time needed to create the mediated schemas.

Figure 3 shows how the execution times of IFD and UDI increase with increasing n up to 17 (the total number of sources in the tested dataset). The impact of the number of sources on the execution time of IFD is not as high as that of UDI. While in the beginning, the execution time of UDI is a little lower than IFD, it dramatically increases eventually. This is because the execution time of IFD is cubic to the number of the attributes of sources (see Section 3.4). But, the execution time of UDI is exponential to the number of uncertain edges. This shows that IFD is much more scalable than UDI.

Effect of FD-point In this section, we study the effect of FD-point on F-measure. Figure 4 shows how F-measure increases with increasing FD-point up

to 680 which is the maximum possible value in the tested dataset. The starting point is when we have one cluster for every attribute. We have not used any recommendation at this point yet; as a result, $FD - point = 0$. Also it is clear that $precision = 1$ and $recall = 0$, thus $F - measure = 0$. As we begin merging clusters using recommendations, FD-point increases and this increases the F-measure as well. The increase in FD-point continues until it reaches its maximum possible value in the tested dataset. We consider all generated mediated schemas with maximum FD-point value as schemas eligible for being in the result set.

5 Related Work

There has been much work in the area of automatic schema matching during the last three decades (see [12] for a survey). They studied how to use various clues to identify the semantics of attributes and match them. An important class of approaches, which are referred to by constraint matchers, uses the constraints in schemas to determine the similarity of schema elements. Examples of such constraints are data types, value ranges, uniqueness, optionality, relationship types, and cardinalities. Our approach is different, since we use an uncertain approach for modeling and generating mediated schemas. Thus, the heuristic rules we use as well as the way we decrease the distance of the attributes is completely different. In addition, we take advantage of FDs. The proposals in [13] and [14] also consider the role of FDs in schema matching. However, our heuristic rules and the way we combine it with attribute similarity is completely different with these proposals.

The closest work to ours is that of Sarma et al. [3] which we denote as UDI in this paper. UDI creates several mediated schemas with probabilities attached to them. To do so, it constructs a weighted graph of source attributes and distinguishes two types of edges: certain and uncertain. Then, a mediated schema is created for every subset of uncertain edges. Our approach has several advantages over UDI. The time complexity of UDI's algorithm for generating mediated schemas is exponential to the number of uncertain edges (i.e. attribute pairs) but that of our algorithm is PTIME (as shown in Section 3.4), therefore our approach is much more scalable. In addition, the quality of mediated schemas generated by our approach has shown to be considerably higher than that of UDI. Furthermore, the mediated schemas generated by our approach are consistent with all sources, while those of UDI may be inconsistent with some sources.

6 Conclusion

In this paper, we proposed IFD, a data integration system with the objective of automatically setting up a data integration application. We established an advanced starting point for pay-as-you-go data integration systems. IFD takes advantage of the background knowledge implied in FDs for finding attribute correlations and using it for matching the source schemas and generating the

mediated schema. We built IFD on a probabilistic data model in order to model the uncertainty in data integration systems.

We validated the performance of IFD through implementation. We showed that using FDs can significantly improve the quality of schema matching (by 26%). We also showed the considerable contribution of using a probabilistic approach (10%). Furthermore, we showed that IFD outperforms UDI, its main competitor, by 23% and has cubic scale up compared to UDI's exponential execution cost.

References

1. Özsu, M.T., Valduriez, P.: Principles of Distributed Database Systems, 3rd Edition. Springer (2011)
2. Madhavan, J., Cohen, S., Dong, X.L., Halevy, A.Y., Jeffery, S.R., Ko, D., Yu, C.: Web-scale data integration: You can afford to pay as you go. In: Proc. of CIDR. (2007)
3. Sarma, A.D., Dong, X., Halevy, A.Y.: Bootstrapping pay-as-you-go data integration systems. In: Proc. of SIGMOD. (2008)
4. Wang, D.Z., Dong, X.L., Sarma, A.D., Franklin, M.J., Halevy, A.Y.: Functional dependency generation and applications in pay-as-you-go data integration systems. In: Proc. of WebDB. (2009)
5. Akbarinia, R., Valduriez, P., Verger, G.: Efficient Evaluation of SUM Queries Over Probabilistic Data. TKDE **to appear** (2012)
6. Dong, X.L., Halevy, A.Y., Yu, C.: Data integration with uncertainty. VLDB J. **18**(2) (2009) 469–500
7. Huhtala, Y., Kärkkäinen, J., Porkka, P., Toivonen, H.: Tane: An efficient algorithm for discovering functional and approximate dependencies. Comput. J. **42**(2) (1999) 100–111
8. Ayat, N., Afsarmanesh, H., Akbarinia, R., Valduriez, P.: Uncertain data integration using functional dependencies. Technical report, In: http://www.science.uva.nl/CO-IM/papers/IFD/ifd.pdf
9. Davidson, I., Ravi, S.S.: Using instance-level constraints in agglomerative hierarchical clustering: theoretical and empirical results. Data Min. Knowl. Discov. **18**(2) (2009) 257–282
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
11. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: Proc. of IIWeb. (2003)
12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB J. **10**(4) (2001) 334–350
13. Biskup, J., Embley, D.W.: Extracting information from heterogeneous information sources using ontologically specified target views. Inf. Syst. **28**(3) (2003) 169–212
14. Larson, J.A., Navathe, S.B., Elmasri, R.: A theory of attribute equivalence in databases with application to schema integration. IEEE Trans. Software Eng. **15**(4) (1989) 449–463