Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching

Sergey Melnik*   Hector Garcia-Molina
Stanford University CA 94305
{melnik,hector}@db.stanford.edu

Erhard Rahm
University of Leipzig, Germany
rahm@informatik.uni-leipzig.de

Abstract

Matching elements of two data schemas or two data instances plays a key role in data warehousing, e-business, or even biochemical applications. In this paper we present a matching algorithm based on a fixpoint computation that is usable across different scenarios. The algorithm takes two graphs (schemas, catalogs, or other data structures) as input, and produces as output a mapping between corresponding nodes of the graphs. Depending on the matching goal, a subset of the mapping is chosen using filters. After our algorithm runs, we expect a human to check and if necessary adjust the results. As a matter of fact, we evaluate the 'accuracy' of the algorithm by counting the number of needed adjustments. We conducted a user study, in which our accuracy metric was used to estimate the labor savings that the users could obtain by utilizing our algorithm to obtain an initial matching. Finally, we illustrate how our matching algorithm is deployed as one of several high-level operators in an implemented testbed for managing information models and mappings.

1. Introduction

Finding correspondences between elements of data schemas or data instances is required in many application scenarios. This task is often referred to as *matching*. Consider a comparison shopping website that aggregates product offers from multiple independent online stores. The comparison site developers need to match the product catalogs of each store against their combined catalog. For instance, the 'product code' field in one catalog may match the 'product ID' and 'store ID' fields in the combined catalog. Or, think of a merger between two corporations, both of which need to consolidate their relational databases deployed by different departments. In this integration scenario, and in many data warehousing applications, match-

---

*On leave from the University of Leipzig

ing of relational schemas is required. Schema matching is utilized for a variety of other types of schemas including UML class taxonomies, ER diagrams, and ontologies.

Matching of data instances is another important task. For example, consider two CAD files or program scripts that have been independently modified by several developers. In this scenario, matching helps identify moved or modified elements in these complex data structures. In bioinformatics, matching has been used for network analysis of molecular interactions [9]. In this domain, data instances represent metabolic networks of chemical compounds, or molecular assembly maps. Matching of molecular networks and biochemical pathways helps to predict metabolism of an organism given its genome sequence.

Matching problems often differ a lot. So do the approaches to matching. For example, for matching relational schemas one could use SQL data types to determine which columns are closely related. On the other hand, in XML schema matching, hierarchical relationships between schema elements can be exploited. Because of this diversity, applications that rely on matching are often built from scratch and require significant amount of thought and programming. We address this problem by proposing a matching algorithm that allows quick development of matchers for a broad spectrum of different scenarios. We are not trying to outperform custom matchers that use highly tuned, domain-specific heuristics.

In this paper we suggest a simple structural algorithm that can be used for matching of diverse data structures. Such data structures, which we call *models*, may be data schemas, data instances, or a combination of both. The elements of models represent artifacts like relational tables and columns, or products and customers. The algorithm that we suggest is based on the following idea. First, we convert the models to be matched into directed labeled graphs. These graphs are used in an iterative fixpoint computation whose results tell us what nodes in one graph are similar to nodes in the second graph. For computing the similarities, we rely on the intuition that elements of two distinct models are

similar when their adjacent elements are similar. In other words, a part of the similarity of two elements propagates to their respective neighbors. The spreading of similarities in the matched models is reminiscent to the way how IP packets flood the network in broadcast communication. For this reason, we call our algorithm the *similarity flooding* algorithm. We refer to the result produced by the algorithm as a *mapping*. Depending on the particular matching goal, we then choose a subset of the resulting mapping using adequate filters. After our algorithm runs, we expect a human to check and if necessary adjust the results. As a matter of fact, in Section 5 we evaluate the 'accuracy' of the algorithm by counting the number of needed adjustments.

While this paper focuses on matching, the broader goal of our work is to design a generic tool that helps to manipulate and maintain schemas, instances, and match results. With this tool, matching is not done entirely automatically. Instead, the tool assists human developers in matching by suggesting plausible match candidates for the elements of a schema. Using a graphical interface, the user adjusts the proposed match result by removing or adding lines connecting the elements of two schemas. Often, the correct match depends on the information only available or understandable by humans. For example, even matches as plausible as ZipCode to zip_code can be doomed as incorrect by a data warehouse designer who knows that zip codes from a given relational source should not be collected due to poor data quality. In such cases, the mappings suggested by the tool may be incorrect or incomplete.

In this paper we are making the following contributions:

- We introduce a generic matching algorithm that is usable across application areas (Section 3).

- We discuss approaches for selecting relevant subsets of match results (Section 4).

- We suggest an 'accuracy' metric for evaluating automatic matching algorithms (Section 5) and evaluate the effectiveness of our algorithm on the basis of a user study that we conducted (Section 6).

We review the related work in Section 7.

2. Overview of the Approach

Before we go into details of our matching algorithm, let us briefly walk through an example that illustrates matching of two relational database schemas. Please keep in mind that the technique we describe is not limited to relational schemas. Consider schemas S_1 and S_2 depicted in Figure 1. The elements of S_1 and S_2 are tables and columns. Assume for now that our goal is to obtain exactly one matching element for every element in S_1. A part of the matching result could be, for example, the correspondence of column
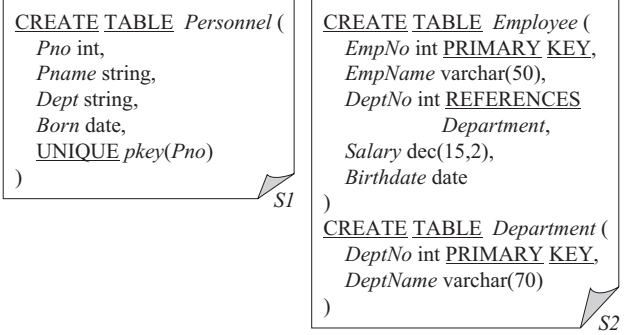

Figure 1. Matching two relational schemas: Personnel and Employee-Department

Personnel/Pname to column Employee/EmpName. A sequence of steps that allows us to determine the correspondences between tables and columns in S_1 and S_2 can be expressed as the following script:

1. G_1 = SQL2Graph(S_1); G_2 = SQL2Graph(S_2);
2. initialMap = StringMatch(G_1, G_2);
3. product = SFJoin(G_1, G_2, initialMap);
4. result = SelectThreshold(product);

As a first step, we translate the schemas from their native format into graphs G_1 and G_2. In our example, the native format of the schemas are ASCII files containing table definitions in SQL DDL. A portion of the graph G_1 is depicted in Figure 2. The translation into graphs is done using an import filter **SQL2Graph** that understands the definitions of relational schemas. We do not insist on choosing a particular graph representation for relational schemas. The representation used in Figure 2 is based on the Open Information Model (OIM) specification [1]. The nodes in the graph are shown as ovals and rectangles. The labels inside the ovals denote the identifiers of the nodes, whereas rectangles represent literals, or string values. For example, node &1 represents the table Personnel in graph G_1, whereas nodes &2, &4, and &6 denote columns Pno, Pname, and Dept, respectively. Column Born and unique key perskey are omitted from the figure for clarity. Tables Employee and Department from schema S_2 are represented in a similar manner in graph G_2. In our example, G_1 has a total of 31 nodes while G_2 has 55 nodes.

As a second step, we obtain an initial mapping initialMap between G_1 and G_2 using operator **StringMatch**. The mapping initialMap is obtained using a simple string matcher that compares common prefixes and suffixes of literals. A portion of the initial mapping is shown in Table 1. Literal nodes are highlighted using apostrophes. The second column of the table lists similarity values between nodes in G_1 and G_2 computed on the basis of their tex-
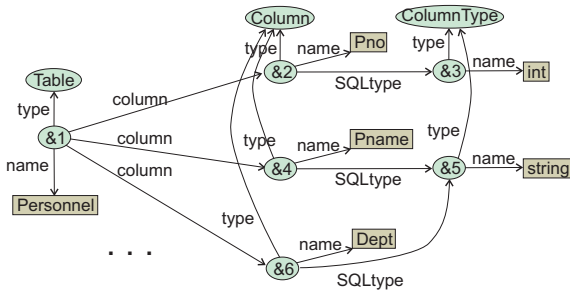
Figure 2. A portion of graph representation G_1 for relational schema S_1

Line#	Similarity	Node in G_1	Node in G_2
1.	1.0	Column	Column
2.	0.66	ColumnType	Column
3.	0.66	'Dept'	'DeptNo'
4.	0.66	'Dept'	'DeptName'
5.	0.5	UniqueKey	PrimaryKey
6.	0.26	'Pname'	'DeptName'
7.	0.26	'Pname'	'EmpName'
8.	0.22	'date'	'Birthdate'
9.	0.11	'Dept'	'Department'
10.	0.06	'int'	'Department'

Table 1. A portion of initialMap **(10 of total 26 entries are shown)**

Similarity	Node in G_1	Node in G_2
1.0	Column	Column
0.81	[Table: Personnel]	[Table: Employee]
0.66	ColumnType	ColumnType
0.44	[ColumnType: int]	[ColumnType: int]
0.43	Table	Table
0.35	[ColumnType: date]	[ColumnType: date]
0.29	[UniqueKey: perskey]	[PrimaryKey: on EmpNo]
0.28	[Col: Personnel/Dept]	[Col: Department/DeptName]
0.25	[Col: Personnel/Pno]	[Col: Employee/EmpNo]
0.19	UniqueKey	PrimaryKey
0.18	[Col: Personnel/Pname]	[Col: Employee/EmpName]
0.17	[Col: Personnel/Born]	[Col: Employee/Birthdate]

Table 2. The mapping after applying Select-Threshold **on result of** SFJoin

tual content. The similarity values range between 0 and 1 and indicate how well the corresponding nodes in G_1 match their counterparts in G_2. Notice that the initial mapping is still quite imprecise. For instance, it suggests mapping column names onto table names (e.g. column Dept in S_1 onto table Department in S_2, line 9), or names of data types onto column names (e.g. SQL type date in S_1 onto column Birthdate in S_2, line 8).

As a third step, operator SFJoin is applied to produce a refined mapping called product between G_1 and G_2. In this paper we propose an iterative 'similarity flooding' (SF) algorithm based on a fixpoint computation that is used for implementing operator SFJoin. The SF algorithm has no knowledge of node and edge semantics. As a starting point for the fixpoint computation the algorithm uses an initial mapping like initialMap. Our algorithm is based on the assumption that whenever any two elements in models G_1 and G_2 are found to be similar, the similarity of their adjacent elements increases. Thus, over a number of iterations, the initial similarity of any two nodes propagates through the graphs. For example, in the first iteration the initial textual similarity of strings 'Pname' and 'EmpName' adds to the similarity of columns Personnel/Pname and Employee/EmpName. In the next iteration, the similarity of Personnel/Pname to Em-

ployee/EmpName propagates to the SQL types string and varchar(50). This subsequently causes increase in similarity between literals 'string' and 'varchar', leading to a higher resemblance of Personnel/Dept to Department/DeptName than that of Personnel/Dept to Department/DeptNo. The algorithm terminates after a fixpoint has been reached, i.e. the similarities of all model elements stabilize. In our example, the refined mapping product returned by SFJoin contains 211 node pairs with positive similarities (out of a total of $31 \cdot 55 = 1705$ entries in the G_1, G_2 cross-product).

As a last operation in the script, operator SelectThreshold selects a subset of node pairs in product that corresponds to the 'most plausible' matching entries. We discuss this operator in Section 4. The complete mapping returned by SelectThreshold contains 12 entries and is listed in Table 2. For readability, we substituted numeric node identifiers by the descriptions of the objects they represent. For example, we replaced node identifier &2 by [Col:Personnel/Pno].

As we see in Table 2, the SF algorithm was able to produce a good mapping between S_1 and S_2 without any built-in knowledge about SQL DDL by merely using graph structures. For example, table Personnel was matched to table Employee despite the lack of textual similarity. Notice that the table still contains correspondences like the one between node Column in G_1 to node Column in G_2, which are hardly of use given our goal of matching the specific tables and columns. We discuss the filtering of match results in more detail in Section 4. The similarity values shown in the table may appear relatively low. As we will explain, in presence of multiple match candidates for a given model element, relative similarities are often more important than absolute values.
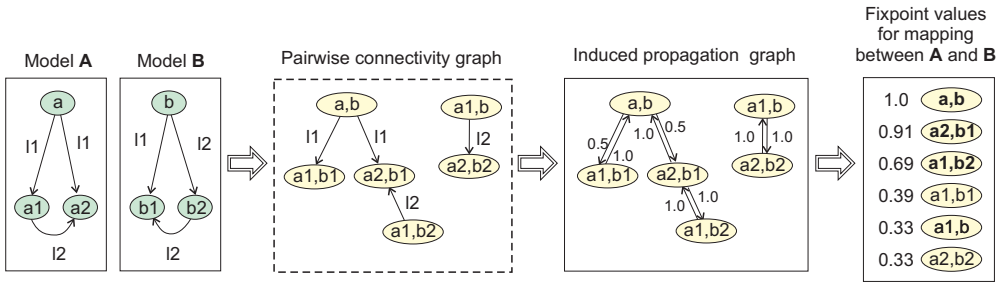
Figure 3. Example illustrating the Similarity Flooding Algorithm

3. Similarity Flooding Algorithm

The internal data model that we use for models and mappings is based on directed labeled graphs. Every edge in a graph is represented as a triple (s, p, o), where s and o are the source and target nodes of the edge, and the middle element p is the label of the edge. For a more formal definition of our internal data model please refer to [14]. In this section, we explain our algorithm using a simple example presented in Figure 3. The top left part of the figure shows two models A and B that we want to match.

Similarity propagation graph A *similarity propagation graph* is an auxiliary data structure derived from models A and B that is used in the fixpoint computation of our algorithm. To illustrate how the propagation graph is computed from A and B, we first define a *pairwise connectivity graph* (PCG) as follows: $((x, y), p, (x', y')) \in \mathrm{PCG}(A, B) \iff (x, p, x') \in A$ and $(y, p, y') \in B$.

Each node in the connectivity graph is an element from $A \times B$. We call such nodes *map pairs*. The connectivity graph for our example is enclosed in a dashed frame in Figure 3. The intuition behind arcs that connect map pairs is the following. Consider for example map pairs (a, b) and (a_1, b_1). If a is similar to b, then probably a_1 is somewhat similar to b_1. The evidence for this conclusion is provided by the l_1-edges that connect a to a_1 in graph A and b to b_1 in graph B. This evidence is captured in the connectivity graph as an l_1-edge leading from (a, b) to (a_1, b_1). We call (a_1, b_1) and (a, b) *neighbors*.

The induced propagation graph for A and B is shown next to the connectivity graph in Figure 3. For every edge in the connectivity graph, the propagation graph contains an additional edge going in the opposite direction. The weights placed on the edges of the propagation graph indicate how well the similarity of a given map pair propagates to its neighbors and back. These so-called *propagation coefficients* range from 0 to 1 inclusively and can be computed in many different ways. The approach illustrated in Figure 3 is based on the intuition that each edge type makes an equal contribution of 1.0 to spreading of similarities from a

given map pair. For example, there is exactly one l_2-edge out of (a_1, b) in the connectivity graph. In such case we set the coefficient $w((a_1, b), (a_2, b_2))$ in the propagation graph to 1.0. The value 1.0 indicates that the similarity of a_1 to b contributes fully to that of a_2 and b_2. Analogously, the propagation coefficient $w((a_2, b_2), (a_1, b))$ on the reverse edge is also set to 1.0, since there is exactly one incoming l_2-edge for (a_2, b_2). In contrast, two l_1-edges are leaving map pair (a, b) in the connectivity graph. Thus, the weight of 1.0 is distributed equally among $w((a, b), (a_1, b_1)) = 0.5$ and $w((a, b), (a_2, b_1)) = 0.5$. In [14] we analyze several alternative ways of computing the propagation coefficients.

Fixpoint computation Let $\sigma(x, y) \geq 0$ be the similarity measure of nodes $x \in A$ and $y \in B$ defined as a total function over $A \times B$. We refer to σ as a mapping. The similarity flooding algorithm is based on an iterative computation of σ-values. Let σ^i denote the mapping between A and B after i^{th} iteration. Mapping σ^0 represents the initial similarity between nodes of A and B, which is obtained e.g., using string comparisons of node labels. In our example we assume that no initial mapping between A and B is available, i.e. $\sigma^0(x, y) = 1.0$ for all $(x, y) \in A \times B$.

In every iteration, the σ-values for a map pair (x, y) are incremented by the σ-values of its neighbor pairs in the propagation graph multiplied by the propagation coefficients on the edges going from the neighbor pairs to (x, y). For example, after the first iteration $\sigma^1(a_1, b_1) = \sigma^0(a_1, b_1) + \sigma^0(a, b) \cdot 0.5 = 1.5$. Analogously, $\sigma^1(a, b) = \sigma^0(a, b) + \sigma^0(a_1, b_1) \cdot 1.0 + \sigma^0(a_2, b_1) \cdot 1.0 = 3.0$. Then, all values are normalized, i.e., divided by the maximal σ-value (of current iteration) $\sigma^1(a, b) = 3.0$. Thus, after normalization we get $\sigma^1(a, b) = 1.0$, $\sigma^1(a_1, b_1) = \frac{1.5}{3.0} = 0.5$, etc. In general, mapping σ^{i+1} is computed from mapping σ^i as follows (normalization is omitted for clarity):

$$\sigma^{i+1}(x, y) = \sigma^i(x, y) + \sum_{(a_u, p, x) \in A, \, (b_u, p, y) \in B} \sigma^i(a_u, b_u) \cdot w((a_u, b_u), (x, y)) + \sum_{(x, p, a_v) \in A, \, (y, p, b_v) \in B} \sigma^i(a_v, b_v) \cdot w((a_v, b_v), (x, y))$$

4

| Identifier	Fixpoint formula
Basic	$\sigma^{i+1} = normalize(\sigma^i + \varphi(\sigma^i))$
A	$\sigma^{i+1} = normalize(\sigma^0 + \varphi(\sigma^i))$
B	$\sigma^{i+1} = normalize(\varphi(\sigma^0 + \sigma^i))$
C	$\sigma^{i+1} = normalize(\sigma^0 + \sigma^i + \varphi(\sigma^0 + \sigma^i))$

Table 3. Variations of the fixpoint formula

The above computation is performed iteratively until the Euclidean length of the residual vector $\Delta(\sigma^n, \sigma^{n-1})$ becomes less than ε for some $n > 0$. If the computation does not converge, we terminate it after some maximal number of iterations. In Section 6 we study the convergence properties of the algorithm. The right part of Figure 3 displays the similarity values for the map pairs in the propagation graph. These values have been obtained after five iterations using the above equation. In the figure, the top three matches with the highest ranks are highlighted in bold. These map pairs indicate how the nodes in A should be mapped onto nodes in B.

Taking normalization into account, we can rewrite the above equation to obtain the 'basic' fixpoint formula shown in Table 3. The function φ increments the similarities of each map pair based on similarities of their neighbors in the propagation graph. The variations A, B, and C of the fixpoint formula are studied in Section 6. Our experiments suggest that formula C performs best with respect to quality of match results and convergence speed.

4. Filters

In this section we examine several filters that can be used for choosing the best match candidates from the list of ranked map pairs returned by the similarity flooding algorithm. Usually, for every element in the matched models, the algorithm delivers a large set of match candidates. Hence, the immediate result of the fixpoint computation (like the one shown on the right of Figure 3) may still be too voluminous for many matching tasks. For instance, in a schema matching application the choice presented to a human user for every schema element may be overwhelming, even when the presented match candidates are ordered by rank. We refer to the immediate result of the iterative computation as *multimapping*, since it contains many potentially useful mappings as subsets.

It is not evident, which criteria could be useful for selecting a desirable subset from a multimapping. An additional complication is that as many as 2^n different subsets can be formed from a set of n map pairs. To illustrate the selection problem, consider the match result obtained for two tiny models A and B that is shown on the left in Figure 4 (the models themselves are are omitted in the figure for clarity). The multimapping M contains four map pairs with similar-
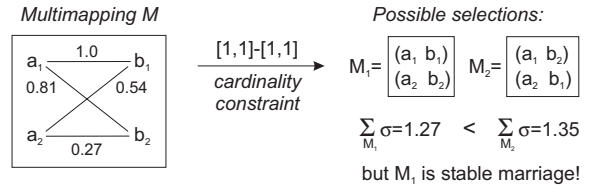
Figure 4. Cumulative similarity vs. 'stable marriage'
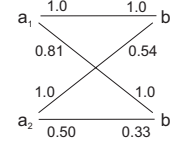
Figure 5. Relative similarities for the example in Figure 4

ities $\sigma(a_1, b_1) = 1.0$, $\sigma(a_2, b_1) = 0.54$, etc. From the set of 4 pairs, $2^4 = 16$ distinct subsets can be selected. Every one of these 16 subsets may be a plausible alternative for the final match result presented to the user.

We address the selection problem using a three-step approach. First, we use the available application-specific constraints to reduce the size of the multimapping. As exemplified below, typing and cardinality constraints may help to eliminate many map pairs from the multimapping. As a second step, we use selection techniques developed in context of matching in bipartite graphs to pick out the subset that is finally delivered to the user. At last, we evaluate the usefulness of particular selection techniques for a given class of matching tasks (e.g. schema matching) and choose the technique with empirically best results. In the rest of this section we discuss the first two steps in more detail. We present an evaluation of several selection techniques in Section 6.

Constraints Frequently, matching tasks include application-specific constraints that can be used for pruning of a large portion of possible selections. Recall our relational schemas S_1 (Personnel) and S_2 (Employee) from Section 2. At least two useful constraints are conceivable for this matching scenario. First, we could use a *typing* constraint to restrict the result to only those matches that hold between columns or tables, i.e., we can ignore matches of keys, data types etc. Second, if our goal were to populate the Personnel table with data from the Employee table, we could deploy a *cardinality* constraint that requires exactly one match candidate for every element of schema S_1. In this case, the cardinality of the resulting mapping would have to satisfy the restriction $[0, n] - [1, 1]$ (using the UML notation). The right expression $[1, 1]$ limits

5

the number of S_2-elements that may match each element of S_1 to exactly one (between a lower limit of 1 and an upper limit of 1). Conversely, the left expression $[0, n]$ specifies the valid number of S_1-match candidates (between 0 and n) for each element of S_2, i.e. elements of S_2 may remain unmatched or may have one or more match candidates.

Unfortunately, in many matching tasks typing or cardinality constraints do not narrow down the match result sufficiently. For example, even after applying a one-to-one ($[1, 1] - [1, 1]$) cardinality constraint in Figure 4, we are still left with two sets of map pairs M_1 and M_2. Below we examine several strategies for making the decision between the remaining alternatives M_i.

Selection metrics   To make an educated choice between M_i's we need an intuition of what constitutes a 'better' mapping. Fortunately, our selection dilemma is closely related to well-known matching problems in bipartite graphs, so that we can build on intuitions and algorithms developed for solving this class of problems (see e.g. [12, 8]). In the graph matching literature, a *matching* is defined as a mapping with cardinality $[0, 1] - [0, 1]$, i.e., a set of edges no two of which are incident on the same node. A *bipartite* graph is one whose nodes form two disjoint parts such that no edge connects any two nodes in the same part. Thus, a mapping can be viewed as an undirected weighted bipartite graph.

A helpful intuition that we will predominantly use for explaining alternative selection strategies for multimappings is provided by the so-called *stable marriage* problem. To remind, in an instance of the stable marriage problem, each of n women and n men lists the members of the opposite sex in order of preference. The goal is to find the best match between men and women. A stable marriage is defined as a complete matching of men and women with the property that there are no two couples (x, y) and (x', y') such that x prefers y' to y and y' prefers x to x'. For obvious reasons, such a situation would be regarded as unstable. Imagine that in Figure 4 elements a_1 and a_2 correspond to women. Then, men b_1 and b_2 would be the primary and the secondary choice for woman a_1. Obviously, mapping M_1 satisfies the stable marriage condition, whereas M_2 does not. In M_2, woman a_1 and man b_1 favor each other over their actual partners, which puts their marriages in jeopardy.

The stable-marriage property provides a plausible criterion for selecting a desired mapping from a multimapping. Further selection criteria can be borrowed from other well-known matching problems like the assignment problem, finding a maximal matching, etc. For example, the assignment problem consists in finding a matching M_i in a weighted bipartite graph M that maximizes the total weight (cumulative similarity) $\sum_{(x,y) \in M_i} \sigma(x, y)$. Viewed as a marriage, such matching maximizes the total satisfaction of all men and women. In Figure 4, $\sum_{M_2} \sigma = 0.81 + 0.54 =$

1.35, whereas $\sum_{M_1} \sigma = 1.0 + 0.27 = 1.27$. Thus, M_2 maximizes the total satisfaction of all men and women even though M_2 is not a stable marriage.

To summarize, the filtering problem can be characterized by providing a set of constraints and a selection function that picks out the 'best' subset of the multimapping under a given selection metric. Conceptually, the selection function assigns a value to every subset of the multimapping. The subset for which the function takes the largest/smallest value is selected as the final result. For example, using the assignment problem as selection metric, we can construct a filter that applies a cardinality constraint $[0, 1] - [0, 1]$ and utilizes a selection function $\sum_{(x,y) \in M_i} \sigma(x, y)$ to choose the best subset. In concrete implementations of selection functions, we can often find algorithms that avoid enumerating all subsets of the multimapping and determine the desired subset directly.

In the remainder of this section we describe a filter that produced empirically best results in a variety of schema matching tasks, as we show later in Section 6. This approach is implemented in our testbed as the **SelectThreshold** operator. The intuition behind this approach is based on a *perfectionist egalitarian polygamy*, which means that no male or female is willing to accept any partner(s) but the best. For a more detailed discussion of alternative selection approaches please refer to [14].

The **SelectThreshold** operator uses *relative* similarities, as opposed to the absolute similarities of map pairs computed by the flooding algorithm. Absolute similarity is symmetric, i.e. x is similar to y exactly as y to x. Under the marriage interpretation, this means that any two prospective partners like each other to the same extent. Considering *relative* similarities suggests a more diversified interpretation. Relative similarities are asymmetric and are computed as fractions of the absolute similarities of the best match candidates for any given element. In the example in Figure 4, b_1 is the best match candidate for a_2, so we set $\overrightarrow{\sigma}_{rel}(a_2, b_1) := 1.0$. The relative similarity for all other match candidates of a_2 is computed as a fraction of $\sigma(a_2, b_1)$. Thus, $\overrightarrow{\sigma}_{rel}(a_2, b_2) := \frac{\sigma(a_2, b_2)}{\sigma(a_2, b_1)} = \frac{0.27}{0.54} = 0.5$. All relative similarities for this example are summarized in Figure 5.

The **SelectThreshold** operator selects a subset of a multimapping, in which all map pairs carry a relative similarity value of at least t_{rel}. For example, for $t_{rel} = 0.5$ in Figure 5, woman a_2 would accept man b_2 as a partner, but man b_2 would reject woman a_2 since $\overleftarrow{\sigma}_{rel}(a_2, b_2) = 0.33 < 0.5$. Most of the time, **SelectThreshold** with $t_{rel} = 1.0$ yields matchings, or monogamous societies. In a less picky version of the operator with $t_{rel} < 1.0$, more persons have a chance to find a partner, and polygamy is more likely. We demonstrate the impact of threshold value t_{rel} in Section 5. **SelectThreshold** operator selects a sub-

set of the multimapping which is guaranteed to satisfy the stable-marriage property in a polygamous society. As we illustrate in Section 6, the stable marriage property proved instrumental for filtering multimappings in schema matching scenarios.

5. Assessment of Matching Quality

In this section, we suggest a metric for measuring the quality of automatic matching algorithms. A crucial issue in evaluating matching algorithms is that a precise definition of the desired match result is often impossible. In many applications the goals of matching depend heavily on the intension of the users, much like the users of an information retrieval system have varying intensions when doing a search. Typically, a user of an information retrieval system is looking for a good, but not necessarily perfect search result, which is generally not known. In contrast, a user performing say schema matching is often able to determine the perfect match result for a given match problem. Moreover, the user is willing to adjust the result manually until the intended match has been established. Thus, we feel that the quality metrics for matching tasks that require tight human quality assessment need to have a slightly different focus than those developed in information retrieval.

The quality metric that we suggest below is based upon *user effort needed to transform a match result obtained automatically into the intended result.* We assume a strict notion of matching quality i.e. being close is not good enough. For example, imagine that a matching algorithm comes up with five equally plausible match candidates for a given element, then decides to return only two of them, and misses the intended candidate(s). In such case, we give the algorithm zero points despite the fact that the two returned candidates might be very similar to what we are looking for. Moreover, our metric does not address semiautomatic matching, in which the user iteratively adjusts the result and invokes repeatedly the matching procedure. Thus, the accuracy results we obtain here can be considered 'pessimistic', i.e., our matching algorithm may be 'more useful' that what our metric predicts.

Matching accuracy Our goal is to estimate how much effort it costs the user to modify the proposed match result $P = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ into the intended result $I = \{(a_1, b_1), \ldots, (a_m, b_m)\}$. The user effort can be measured in terms of additions and deletions of map pairs performed on the proposed match result P. One simplified metric that can be used for this purpose is what we call *match accuracy*. Let $c = \|P \cap I\|$ be the number of correct suggestions. The difference $(n - c)$ denotes the number of false positives to be removed from P, and $(m - c)$ is the number of false negatives, i.e. missing matches that need to

be added. For simplicity, let us assume that deletions and additions of match pairs require the same amount of effort, and that the verification of a correct match pair is free. If the user performs the whole matching procedure manually (and does not make mistakes), m add operations are required. Thus, the portion of the manual clean-up needed after applying the automatic matcher amounts to $\frac{(n-c)+(m-c)}{m}$ of the fully manual matching.

We approximate the labor savings obtained by using an automatic matcher as accuracy of match result, defined as $1 - \frac{(n-c)+(m-c)}{m}$. In a perfect match, $n = m = c$, resulting in accuracy 1. Notice that $\frac{c}{m}$ and $\frac{c}{n}$ correspond to recall and precision of matching [11]. Hence, we can express match accuracy as a function of recall and precision as follows:

$$\begin{aligned} \text{Accuracy} \quad &= 1 - \frac{(n-c)+(m-c)}{m} = \frac{c}{m}\left(2 - \frac{n}{c}\right) \\ &= \text{Recall}\left(2 - \frac{1}{\text{Precision}}\right) \end{aligned}$$

In the above definition, the notion of accuracy only makes sense if precision is not less than 0.5, i.e at least half of the returned matches are correct. Otherwise, the accuracy is negative. Indeed, if more than a half of the matches are wrong, it would take the user more effort to remove the false positives and add the missing matches than to do the matching manually from scratch. As expected, the best accuracy 1.0 is achieved when both precision and recall are equal to 1.0.

Intended match result Accuracy, as well as recall and precision, are relative measures that depend on the *intended match result*. For a meaningful assessment of match quality, the intended match result must be specified precisely. Recall our example dealing with relational schemas that we examined in Section 2. Three plausible match results for this example (that we call Sparse, Expected, and Verbose) are presented in Table 4. A plus sign (+) indicates that the map pair shown on the right is contained in the corresponding desired match result. For example, map pair ([Table: Personnel], [Table: Employee]) belongs to both Expected and Verbose intended results. The Expected result is the one that we consider the most natural one. The Verbose result illustrates a scenario where matches are included due to additional information available to the human designer. For example, the data in table Personnel is obtained from both Employee and Department, although this is not apparent just by looking at the schemas. Similarly, the Sparse result is a matching where some correspondences have been eliminated due to application-dependent semantics. Keep in mind that in the Sparse and Verbose scenarios, the human selecting the 'perfect' matchings has more information available than our matcher. Thus, clearly we cannot expect our matching algorithm to do as well as in the Expected case.

Sparse	Expected	Verbose	Node in G_1	Node in G_2
	+	+	[Table: Personnel]	[Table: Employee]
		+	[Table: Personnel]	[Table: Department]
	+	+	[UniqueKey: perskey]	[PrimaryKey: on EmpNo]
+	+	+	[Col: Personnel/Dept]	[Col: Department/DeptName]
		+	[Col: Personnel/Dept]	[Col: Department/DeptNo]
		+	[Col: Personnel/Dept]	[Col: Employee/DeptNo]
+	+	+	[Col: Personnel/Pno]	[Col: Employee/EmpNo]
+	+	+	[Col: Personnel/Pname]	[Col: Employee/EmpName]
+	+	+	[Col: Personnel/Born]	[Col: Employee/Birthdate]

Table 4. Three plausible intended match results for matching problem in Figure 1
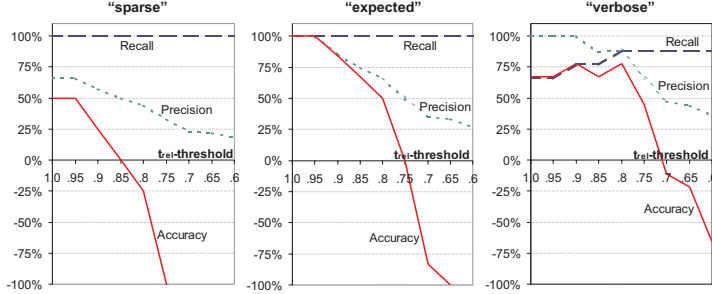

Figure 6. Matching accuracy as a function of t_{rel}-threshold for intended match results Sparse, Expected, and Verbose from Table 4

Accuracy, precision, and recall obtained for all three intended results using version C of the flooding algorithm (see Table 3) are summarized in Figure 6. For each diagram, we executed a script like the one presented in Section 2. The SelectThreshold operator was parameterized using t_{rel}-threshold values ranging from 0.6 to 1.0. As an additional last step in the script, we applied operator SQLMapFilter that eliminates all matches except those between tables, columns, and keys.[1] As shown in the figure, match accuracy 1.0 is achieved for $0.95 \leq t_{rel} \leq 1.0$ in the Expected match, i.e., no manual adjustment of the result is required from the user. In contrast, if the intended result is Sparse, the user can save only 50% of work at best. Notice that the accuracy quickly becomes negative (precision goes below 0.5) with decreasing threshold values. Using no threshold filter at all, i.e. $t_{rel} = 0$, yields recall of 100% but only 4% precision, and results in a desastrous accuracy value of –2144% (not shown in the figure). Increasing threshold values corresponds to the attempt of the user to quickly prune undesired results by adjusting a threshold slider in a graphical tool.

Figure 6 indicates that the quality of matching algorithms may vary significantly in presence of different matching goals. As mentioned earlier, our definition of accuracy is pessimistic, i.e., the user may save more work as indicated by the accuracy values. The reason for that is twofold. On the one hand, if accuracy goes far below zero, the user will probably scrap the proposed result altogether and start from scratch. In this case, no additional work (in contrast to that implied by negative accuracy) is required. On the other hand, removing false positives is typically less labor-intensive than finding the missing match candidates. For example, consider the data point $t_{rel} = 0.75$ in the Expected diagram. The matcher found all 6 intended map pairs (100% recall), and additionally returned 6 false positives (50% precision) resulting in an accuracy of 0.0. Arguably, removing these false positives requires less work as compared to starting with a blank screen.

6. Evaluation of algorithm and filters

To evaluate the performace of the algorithm for schema matching tasks, we conducted a user study with help of eight volunteers in the Stanford Database Group. The study also helped us to examine how different filters and parameters of the algorithm affect the match results. For our study we used nine relatively simple match problems.[2] Some of the problems were borrowed from reseach papers [15, 6, 18]. Others were derived from data used on the web-

---

[1]SQLMapFilter does not filter out (unlikely) matches between say tables and columns.

[2]The complete specification of the match tasks handed out to the users is available at http://www-db.stanford.edu/~melnik/mm/sfa/
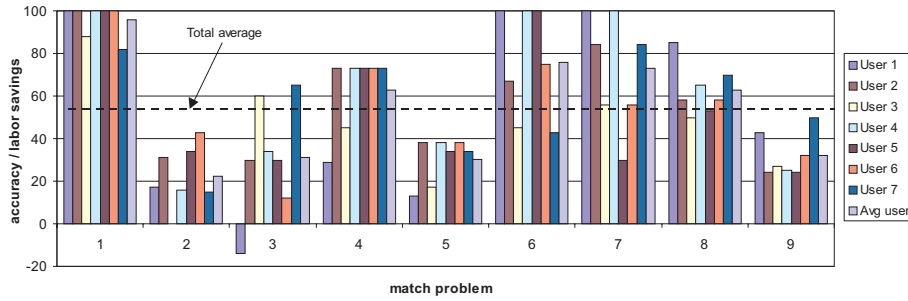
Figure 7. Average matching accuracy for 7 users and 9 matching problems

sites like Amazon.com or Yahoo.com. Every user was required to solve tasks of three different kinds (shown along the x-axis of Figure 7):

1. matching of XML schemas (Tasks 1,2,3)

2. matching of XML schemas using XML data instances (Tasks 4,5,6)

3. matching of relational schemas (Tasks 7,8,9)

The information provided about the source and target schemas was intentionally vague. The users were asked to imagine a plausible scenario and to map elements in both schemas according to the scenario they had in mind. No cardinality constraints were given (any $[0, n] - [0, n]$ mapping was accepted). Noteworthy is that almost no two users could agree on the intended match result for a given matching task, even when examples of data instances were provided (tasks 4,5,6). Therefore, we could hardly expect any automatic procedure to produce excellent results. From eight users, one outlier (i.e. the user with highly deviating results) was eliminated. The accuracy in percent achieved by our algorithm (using fixpoint formula C) for each of the seven users and every task is summarized in Figure 7. The accuracy metric was used to estimate the amount of work that a given user could save by using our algorithm. The accuracy data was obtained after applying SelectThreshold operator with $t_{rel} = 1$. Negative accuracy of –14% in Task 3 indicates that User 1 would have spent 14% more work adjusting the automatic match result than doing the match manually.

Note that in Task 1 the algorithm performed very well, while in Task 2 the results were poor. It turned out that the models used in Task 2 had very simple structure, so that the algorithm was mainly driven by the initial textual match. We did not use any dictionaries for string matching in any of the experiments reported in this paper. Hence, the synonyms used in Task 2 were considered as plausible matches by humans but were not recognized by the algorithm. The matching accuracy over 7 users and 9 problems averaged
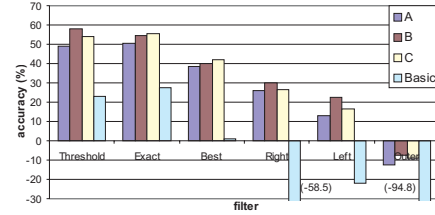

Figure 8. Matching accuracy for different filters and three versions of the algorithm

to 52%. Hence, our study suggests that for many matching tasks, as much as a half of manual work can be saved using very little application-specific code. This figure is typically even higher in simpler tasks, e.g. when matching two XML documents conforming to the same DTD. Using synonyms may further improve the results of matching. The sizes of the propagation graphs obtained from schemas used in the study ranged from 128 to 1222 edges.

Using matching accuracy as the quality measure, we utilized the data collected in the user study to drive our evaluation and tuning of the algorithm for schema matching. As a result of this evaluation, we determined the parameters of the algorithm and the filter that performed best on average for all users and matching problems in our study. The variations of the fixpoint formula that we used are depicted in Table 3. Using distinct fixpoint formulas results in different multimappings produced by the algorithm as well as different convergence speed. We then applied different filters to choose the best subsets of multimappings. Figure 8 summarizes the accuracy (averaged over all tasks) obtained for every version of the algorithm and filter that we used. The filters were defined as follows:

- *Threshold* filter corresponds to the SelectThreshold operator described in Section 4. It produces mappings of cardinality $[0, n] - [0, n]$ using relative-similarity threshold $t_{rel} = 1.0$.

9

- *Exact* is a $[0, 1] - [0, 1]$ version of *Threshold*, which yields monogamous societies.

- *Best* returns a $[0, 1] - [0, 1]$ mapping using a selection metric that corresponds to the assignment problem. The implementation of the filter uses a greedy heuristic. For the next unmatched element, a best available candidate is chosen that maximizes the cumulative similarity.

- *Left* yields a $[0, 1] - [1, 1]$ mapping, in which every node on the left is assigned a match candidate that maximizes the cumulative similarity. *Right* is a $[1, 1] - [0, 1]$ counterpart of *Left*.

- *Outer* filter delivers a $[1, n] - [1, n]$ mapping, in which every node on the left and on the right is guaranteed to have at least one match candidate.

As suggested by Figure 8, the best overall accuracy of 57.9% was achieved using *Threshold* filter with the fixpoint formula B. The accuracy of *Threshold* and *Exact* filters lie very close to each other. This is not surprising, since *Threshold* with $t_{rel} = 1.0$ typically produces $[0, 1] - [0, 1]$ mappings. In our study, *Right* consistently outperforms *Left*, since in most matching tasks the right schemas were smaller; nodes in right schemas were therefore more likely to appear in the intended match results supplied by the users. *Outer* performed worst, since in many tasks only small portions of schemas were intended to have matching counterparts.

We tried to estimate the usefulness of other filters, which are either hard to implement or require extensive computation, by using sampling. For example, a filter that returns a maximal matching (a $[0, 1] - [0, 1]$ mapping with the most map pairs) is apparently not an optimal one for schema matching. Under formula B, the total number of map pairs in all tasks after applying the *Best* filter is 101, with associated accuracy of 40%. This accuracy value is lower than 54% obtained using the *Exact* filter that yields only 73 map pairs. Overall, our study suggests that preserving the stable-marriage property is desirable for selecting subsets of multimappings.

Notice that the fixpoint formulae A, B, and C yield comparable matching accuracy for each filter. However, formula C has much better convergence properties, as suggested by Table 5. The table shows the number n of iterations that were required in every task to obtain a residual vector $\|\Delta(\sigma^n, \sigma^{n-1})\| < 0.05$. For every fixpoint formula, we executed the algorithm in two versions, 'as is' and 'strongly connected'. Strongly connected versions guarantee convergence. This effect is achieved by making σ^0 contain positive similarity values (e.g. at least 0.001) for each map pair in the cross-product of nodes of left and right
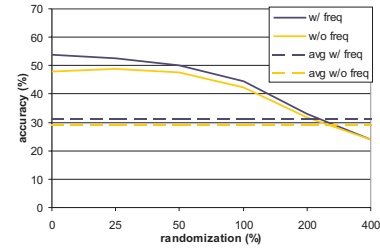
Figure 9. Impact of randomizing initial similarities on matching accuracy

schemas. We found experimentally that the strongly connected versions of the algorithm yielded approximately the same overall accuracy for the filters that preserve the stable-marriage property (*Threshold*, *Exact*, and *Best*). In contrast, enforcing convergence had a substantial negative impact on accuracy for the filters *Left*, *Right*, and *Outer*. For a detailed dicussion of convergence criteria please refer to [14].

We investigated seven distinct approaches to computing the propagation coefficients in the induced propagation graph. In the approach illustrated in Section 3, we use a so-called *inverse-product* formula: we count the number of arcs with a certain label that leave a given pair of nodes in both graphs, multiply the two numbers, and take the inverse of this product as a propagation coefficient. In our user study, we found that the best overall match results were produced using the inverse-average formula. In this formula, we take the inverse of the *average* number (instead of product) of equilabeled arcs. A formal definition of the formulas that we examined and the details of this experiment are presented in [14].

As a last experiment in this section, we study the impact of the initial similarity values (σ^0) on the performance of the algorithm. For this purpose, we randomly distorted the initial values computed by the string matcher. The initial similarities were computed using two versions of a string matcher, one of which took term frequencies into account. Figure 9 depicts the influence of randomization on matching accuracy across all users and matching tasks. For example, randomization of 50% means that every initial similarity value was randomly increased or decreased by x percent, $x \in [-50\%, 50\%]$. Negative similarity was adjusted to zero. It is noteworthy that a randomization factor of 100% introduced accuracy penalty of just about 15%. This result indicates that the similarity flooding algorithm is relatively robust against variations in seed similarities. The dotted lines show another radical modification of initial similarities, in which each non-zero value in σ^0 was set to the same number computed as the average of all positive similarity values. In this case, the accuracy dropped to 30%, which

Formula	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Total
A (as is)	18	48	122	78	∞	12	37	25	25	∞
A (strongly connected)	15	56	89	81	1488	18	48	25	31	1851
B (as is)	8	428	17	39	8	13	10	24	21	568
B (strongly connected)	7	268	21	32	13	15	14	21	53	444
C (as is)	7	9	9	11	7	7	9	10	9	78
C (strongly connected)	7	9	8	11	7	5	9	7	9	72

Table 5. Illustration of convergence properties of variations of fixpoint formula for tasks $T_1, \ldots, T_9$ in the user study. Shows iterations needed until length of residual vector < 0.05.

still saves the users on average one third of the manual work.

To summarize, the main results of our study were the following:

- For an average user, overall labor savings across all tasks were above 50%. Recall from Section 5 that our accuracy metric gives a pessimistic estimate, i.e. actual savings may be even higher.

- A quickly converging version of the fixpoint formula (C) did not introduce accuracy penalties.

- *Threshold* filter performed best.

- The best formula for computing the propagation coefficients was based on the inverse average [14].

- The flooding algorithm is relatively insensitive to 'errors' in initial similarity values.

7. Related Work

Our work was inspired by model management scenarios presented in the vision paper [2] by Bernstein et al. In particular, our scripts use similar high-level operations on models. Such an approach can significantly simplify the development of metadata-based tasks and applications compared to the use of current metadata repositories and their low-level APIs.

A recent classification and review of matching techniques can be found in [18]. Most of the previously proposed approaches lack genericity and are tailored to a specific application domain such as schema or data integration, and specific schema types such as relational or XML schemas. Moreover, most approaches are restricted to finding 1:1 matching correspondences. A few promising approaches not only use schema-level but also instance-level information for schema matching [11, 6]. Unfortunately, their use of neural networks [11] or machine learning techniques [6] introduces additional preparing and training effort.

Concurrently and independently to the work reported in this paper, a generic schema matching approach called Cupid was developed at Microsoft Research [13]. It uses a comprehensive name matching based on synonym tables and other thesauri as well as a new structural matching approach considering data types and topological adjacency of schema elements.

Many other studies have used more sophisticated linguistic (name/text) matchers compared to our very simple string matcher, e.g. WHIRL [5]. The work in [15] addresses the related problem of determining mapping expressions between matching elements.

In general, match algorithms developed by different researchers are hard to compare since most of them are not generic but tailored to a specific application domain and schema types. Moreover, as we have discussed in Section 5, matching is a subjective operation and there is not always a unique result. Previously proposed metrics for measuring the matching accuracy [11, 6] did not consider the extra work caused by wrong match proposals. Our accuracy metric is similar in spirit to measuring the length of edit scripts as suggested in [4]. However, we are counting the edit operations on mappings, rather than those performed on models to be matched.

In designing our algorithm and the filters, we borrowed ideas from three research areas. The fixpoint computation corresponds to random walks over graphs [16], as explained in [14]. A well-known example of using fixpoint computation for ranking nodes in graphs is the PageRank algorithm used in the Google search engine [3]. Unlike PageRank, our algorithm has two source graphs and extensively uses and depends on edge labeling. The filters that we proposed for choosing subsets of multimappings are based on the intuition behind the class of stable marriage problems [8]. General matching theory and algorithms are comprehensively covered in [12]. Finally, the quality metric that we use for evaluating the algorithm is related to the precision/recall metrics developed in the context of information retrieval.

The data model used in this paper is based on the RDF model [10]. For transforming native data into graphs we use graph-based models defined for different applications (see e.g. [1, 17, 7]).

8. Conclusion

In this paper we presented a simple structural algorithm based on fixpoint computation that is usable for matching of diverse data structures. We illustrated the applicability of the algorithm to a variety of scenarios. We defined several filtering strategies for pruning the immediate result of the fixpoint computation. We suggested a novel quality metric for evaluating the performance of matching algorithms, and conducted a user study to determine which configuration of the algorithm and filters performs best in chosen schema matching scenarios.

The similarity flooding algorithm and the filters discussed in the paper are used as operators in the testbed that we implemented. In our testbed, high-level algebraic operations are deployed for manipulating models and mappings using scripts like the one shown in Section 1. The testbed supports schema definitions and instance data available in SQL DDL, XML, OEM, UML, and RDF.

Additional aspects of our work are covered in the extended technical report [14]. These aspects include convergence and complexity of the flooding algorithm, a summary of open issues and limitations, architecture and implementation of the testbed, and several detailed examples that illustrate computing schema correspondences using instance data, or finding related elements in a data instance.

Acknowledgements

References

[1] P. A. Bernstein, T. Bergstraesser, J. Carlson, S. Pal, P. Sanders, and D. Shutt. Microsoft Repository Version 2 and the Open Information Model. *Information Systems*, pages 71–98, 1999.

[2] P. A. Bernstein, A. Halevy, and R. Pottinger. A Vision for Management of Complex Models. *SIGMOD Record*, pages 55–63, 2000.

[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. WWW7 Conf.* Computer Networks, 1998.

[4] S. S. Chawathe and H. Garcia-Molina. Meaningful Change Detection in Structured Data. In *Proc. SIGMOD'97*, pages 26–37, 1997.

[5] W. W. Cohen. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity. In *Proc. SIGMOD 1998*, pages 201–212, 1998.

[6] A. Doan, P. Domingos, and A. Halevy. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. In *Proc. SIGMOD 2001*, 2001.

[7] XML Document Object Model (DOM), W3C Recommendation. `http://www.w3.org/TR/REC-DOM-Level-1/`, Oct. 1998.

[8] D. Gusfield and R. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, Cambridge, MA, 1989.

[9] M. Kanehisa. *Post-Genome Informatics*. Oxford University Press, 2000.

[10] O. Lassila and R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. `http://www.w3.org/TR/REC-rdf-syntax/`, 1998.

[11] W.-S. Li and C. Clifton. SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. *Trans. on Data & Knowledge Engineering*, pages 49–84, 2000.

[12] L. Lovàsz and M. Plummer. *Matching Theory*. North-Holland, Amsterdam, 1986.

[13] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic Schema Matching with Cupid. In *Proc. 27th VLDB Conf.*, Sept. 2001.

[14] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity Flooding: A Versatile Graph Matching Algorithm. Extended Technical Report, `http://dbpubs.stanford.edu/pub/2001-25`, 2001.

[15] R. J. Miller, L. M. Haas, and M. A. Hernandez. Schema Mapping as Query Discovery. In *Proc. VLDB 2000*, 2000.

[16] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[17] Y. Papakonstantinou, H. García-Molina, and J. Widom. Object Exchange Across Heterogeneous Information Sources. In *Proc. of the 11th IEEE Int. Conf. on Data Engineering (ICDE)*, pages 251–260, Taipe, Taiwan, Mar. 1995.

[18] E. Rahm and P. A. Bernstein. On Matching Schemas Automatically. Technical Report MSR-TR-2001-17, http://www.research.microsoft.com/pubs/, Feb. 2001.