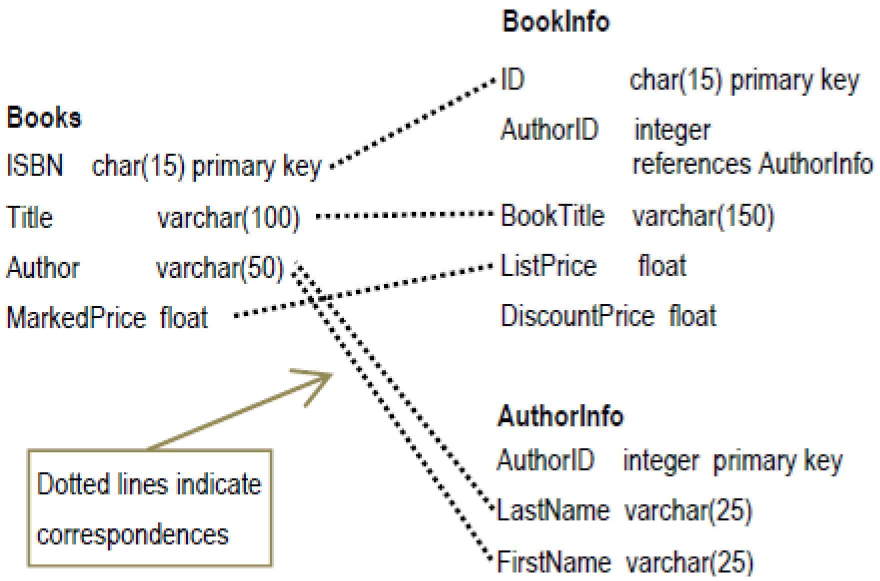
A new dimension in schema matching

杜嘉桐 黄殊凡

Introduction

Schema matching is the problem of discovering the correspondence between two schemas' elements. The term schema refers to the structure in SQL, XML, ontology description, etc. to describe or define the attribute of the ontologies. Correspondence means the relationship between several elements (usually one) of one schema and several elements of another one. In most cases, the correspondence is one-to-one, but sometimes it's one to N or even N to N. If the instances of two related elements are equal, then the one-to-one correspondence can be considered valid. For example, in the figure, the dotted line infers the correspondence between two elements. The special case is the correspondence between author and the pair last name, first name, it's not an one-to-one correspondence but an one-to-two correspondence.



Over the years, research in schema matching has developed extensively. It is impossible to define a mapping expression and find an exact match, or define a definite correspondence set. In addition, in practical applications, the problem of vast amount of data makes manual fitting impossible. Since it is impossible to find an exact match, schema matchers can only perform a best effort matching.

However, there is no indication of the prospective success of the matchers. So we can import the method of schema matching prediction, to assess the performance of schema matchers without an exact match. Predictors predict the performance of the matcher in identifying correct correspondence by analyzing the similarity scores.

The value of prediction is not limited in assessing the matcher. Predictors can also be the basis of decision under uncertainty, by means of assessing the quality of the match. It can help deciding which web form to access data through. Also, the predictor can also serve as a comparative standard in deciding which matcher to use, explaining the outcome of matcher, and showing the improvement direction of matchers. With the predictor, it is possible to change the matcher weights dynamically and tune the matching task automatically, making automated web form

integration possible.
-In this work we propose a new predictor that is based on comparing the strength of a matcher confidence in a pair.

Background

The use of Web data structural features was shown to be beneficial in the context of information seeking on the Web. In some cases, information annotation in the form of, e.g., ontologies, can assist in overcoming the ambiguity inherent to natural languages.
In health care, it may arise in the alignment of patient records and other medical reports. In web applications, it may be used to align product catalogs. In ecommerce, it may be used to align message formats representing business documents, such as orders and invoices.

Related works

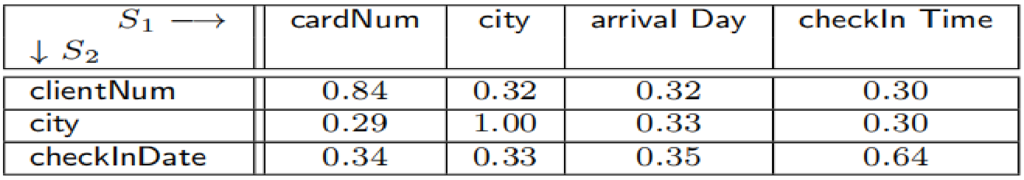
In predicting the quality of the outcome of a schema matching process, the task has been seldom mentioned and often taken as side remarks. In one paper, the concept of similarity space is defined and applied the concept to predictors by presenting a set of predictors and introducing prediction models.

Motivation

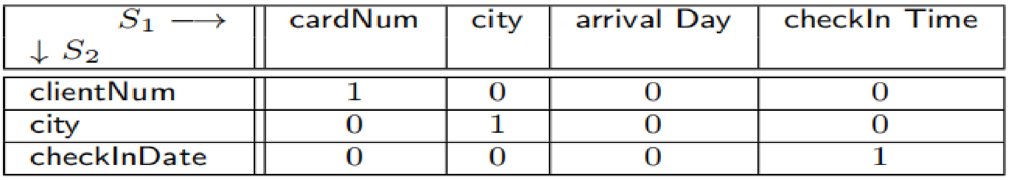
In reading the paper, a problem is raised, which is whether we can find another predictor, other than those presented in the paper, to get a more precise assessment on the performance of schema matchers. Under this consideration, we present a method to define a new dimension of predictors to better predict the result of matchers with independency.
Problem Formulation
Schema matching is the problem of generating correspondences between elements of two schemas. It can be useful when dealing with problems like combining data or transfering models from one dataset to another.
Current research has divided the problem into two steps. The first is to generate a Similarity Space to describe the schemas and the second is to use Predictors to assess the quality of a matching result.
The main problem in this problem is how to maintain the balance between correlation and robustness because the content of schemas varies.
What we want to do is to make alternations to the current methods of generating similarity spaces and predictors to obtain better precision without affecting the robustness.

Proposed method

We now present a model for schema matching. Matching problems match two members of problem domain (schemata) by aligning their components (attributes). Therefore let S_1, S_2 be two schemata $\{a_1, a_{2,} . . . , a_N \}$ and $\{b_1, b_2 , . . . , b_N \}$,
Definition 1. let $S = S_1 \times S_2$, $M(S_1, S_2)$ is an $n \times m$ binary similarity matrix .
For any matched schema pair (S_1,S_2) the power-set $\Sigma=2^s$ is the set of all possible matches between this pair. We denote a match by $\Theta \in \Sigma$ and its cardinality is $| \Theta |$
e.g



$S_1 \longrightarrow$ $\downarrow S_2$	cardNum	city	arrival Day	checkIn Time
clientNum	0.84	0.32	0.32	0.30
city	0.29	1.00	0.33	0.30
checkInDate	0.34	0.33	0.35	0.64

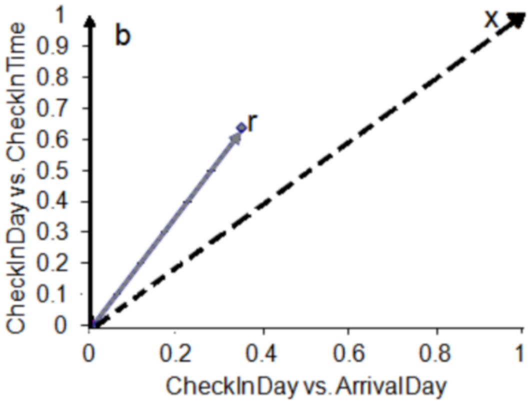


$S_1 \longrightarrow$ $\downarrow S_2$	cardNum	city	arrival Day	checkIn Time
clientNum	1	0	0	0
city	0	1	0	0
checkInDate	0	0	0	1

As an example, consider Table 1, which presents two similarity matrices for two simplified database schemata, with four and three attributes, respectively. We interpret binary similarity matrices as representing a match, where a value of 1 signifies attribute pairs that are part of a match. Therefore, the match that is represented by Table 1(bottom) is σ = h(cardNum, clientNum), (city, city), (checkInTime,checkInDate)i, and its cardinality is | σ | = 3.

Matching is often a stepped process in which different algorithms, rules, and constraints are applied.

We separate matchers into those that are applied directly to the problem (frist-line matchers – 1LMs) and those that are applied to the outcome of other matchers (second-line matchers – 2LMs). 1LMs receive two schemata and return a similarity matrix, like word-vector



2LMs receive a similarity matrix and return a similarity matrix .Using predictor is an improvement of 2LM.

our work is to calculate the independence captured by measuring how much each matrix entry (i, j) ∈ σ, selected by a 2LM, deviates (in terms of match confidence) from other competing entries (i, l); l = j or (l, j); l = i in the similarity matrix M

Schema matching predictors assess the quality of the matching outcome without any knowledge of the exact match. Such prediction can be based on either internal properties of the similarity matrix or by a distance measure from some “ideal” form of solution. Predictors should be applied to tasks with different requirements of granularity,

from predicting match quality for a single attribute pair, to match quality of a schema pair. Predictors should be able to predict different qualities, putting more emphasis, for example, on

Precision or on Recall. Quality of predictors is measured by its correlation with match Precision or Recall, and a good correlation should be statistically significant when tested over a substantial number of schema pairs and stable over varying datasets and schema matchers. Our work is a new predictor,

which measures the diversity of a match $\sigma \in \Sigma$ that was determined by some 2LM, given a similarity matrix M. Informally, match diversity is captured by measuring how much each matrix entry (i, j) $\in \sigma$, selected by a 2LM, deviates (in terms of match confidence) from other competing entries

(i, l); l = j or (l, j); l = i in the similarity matrix M.

More formally, deviation is captured by measuring the difference between entry (i, j) $\in \sigma$ confidence $M_{i,j}$ and that of a mean entry, $\mu_{i,j}$, defined as the average confidence among entries that share the same matrix row i or column j (including entry (i, j)), as follows:

$$\mu_{i,j} = \frac{1}{n+m-1}\left(\sum_{l=1}^{n} M_{l,j} + \sum_{l=1}^{m} M_{i,l} - M_{i,j}\right)$$

$$\Delta_{i,j} = (M_{i,j} - \mu_{i,j})^2$$

We propose an new method to calculate the independence of each i in M and add this to The recently methodFor a given similarity matrix M (generated by some 1LM)

and a match $\sigma \in \Sigma$ (generated by some 2LM), the

predictor evaluates the quality of the match according to

the average (scaled) deviation, as follows:

$$\mathrm{D}(\sigma, M) = \sqrt{\frac{1}{|\sigma|}\sum_{(i,j)\in\sigma} \Delta_{i,j}}$$

Experiment

1LM

Using edit distance and soundex, identify syntactically similar attributes.

WordNet-based algorithm to calculate the similarity matrix M

2LM

We using an 2LM Algorithm MDT which considers that higher valued matrix entries are more likely to be correct than lower ones

$$\delta_i = \text{Delta} \times \left(max_i - \frac{1}{m}\sum_{j=1}^{m} \Delta_{i,j}\right)$$

where MAX(i)=MAX(j)=1······m .Those entries with higher predicted values are selected. The selected for the match Θ holds the following condition

$$\Delta_{i,j} \geq max_i - \delta_i.$$

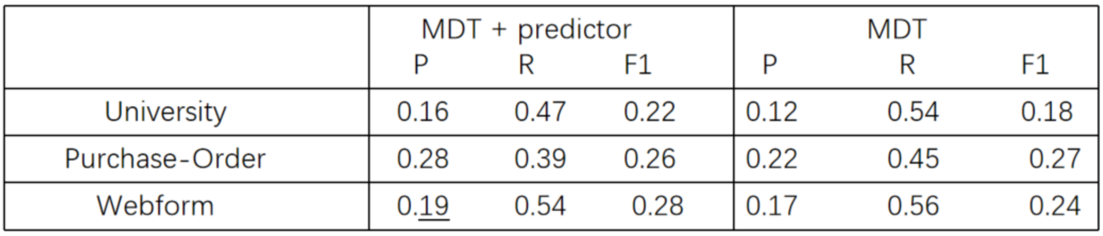
The pseudo code of our algorithm is shown below.

```
input(M(n,m))
for (i,j)∈M do
     Δ （i,j） =M(i,j)- μ (i,j)
end for
k=min(n,m)
for p =1……k do
    Θ =MDT （ Δ ,p）
    if D( Θ ， M) > D( Θ *， M) then
       Θ * = Θ
    end if
end for
return  Θ *
```

And we also compare our work to the existing predictor like STDEV BMM

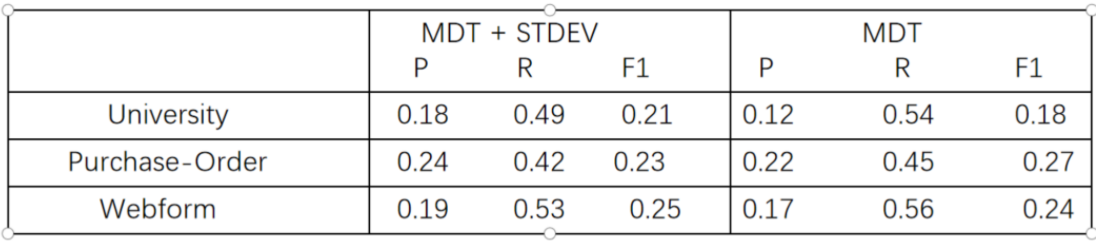
Result

We now compare the independence enhanced MDT against its basic version
Using three 1LMS result which based on the University data-set , Purchase-Order data-set and webform data-set. randomly pick 100 matrix and we set Delta= 0.1 for all MDT Experiments



	MDT + predictor			MDT		
	P	R	F1	P	R	F1
University	0.16	0.47	0.22	0.12	0.54	0.18
Purchase-Order	0.28	0.39	0.26	0.22	0.45	0.27
Webform	0.19	0.54	0.28	0.17	0.56	0.24

We can see that data in University and purchase-Order is well-performed while the Webform
The reason why P and R of this project is quite low is we delete the same match on the result
For e.g:
Suppose we have 100 word pair and 80 percent is exactly the same while (FirstName and first Name) we delete this kind of result and predict the left ones and the method we use MDT is not OPT but we only focus on the improvement of it.
And when compare this predictor to the other one like STDEV



	MDT + STDEV			MDT		
	P	R	F1	P	R	F1
University	0.18	0.49	0.21	0.12	0.54	0.18
Purchase-Order	0.24	0.42	0.23	0.22	0.45	0.27
Webform	0.19	0.53	0.25	0.17	0.56	0.24

Conclusion

In this work we presented a new schema matching predictor, discussed it’s independence and used it to enhance the performance of two existing state-of -the -art schema matchers Our empirical evaluation shows that Our work to be more predictive than other predictors in the literature at some types of data .We also demonstrated empirically its usefulness for schema matching in general. Our work can be extended in several ways First we intend to test the MDT on additional schema matchers. Second an important observation from this work is that diversification in schema matching plays an important role .Hence, we would like to explore additional methods for schema matching diversification and analyze their impact on quality using the evaluation methodology proposed in this work.