Interactive User Feedback in Ontology Matching Using Signature Vectors

Isabel F. Cruz, Cosmin Stroe
ADVIS Lab, Department of Computer Science
University of Illinois at Chicago
USA
{ifc,cstroe1}@cs.uic.edu

Matteo Palmonari
DISCo
University of Milano-Bicocca
Italy
palmonari@disco.unimib.it

***Abstract*— When compared to a *gold standard*, the set of mappings that are generated by an automatic ontology matching process is neither complete nor are the individual mappings always correct. However, given the explosion in the number, size, and complexity of available ontologies, domain experts no longer have the capability to create ontology mappings without considerable effort. We present a solution to this problem that consists of making the ontology matching process interactive so as to incorporate user feedback in the loop. Our approach clusters mappings to identify where user feedback will be most beneficial in reducing the number of user interactions and system iterations. This feedback process has been implemented in the AgreementMaker system and is supported by visual analytic techniques that help users to better understand the matching process. Experimental results using the OAEI benchmarks show the effectiveness of our approach. We will demonstrate how users can interact with the ontology matching process through the AgreementMaker user interface to match real-world ontologies.**

I. Introduction

The ontology matching problem, which is related to schema matching in databases [1], consists of mapping concepts in a *source* ontology to semantically related concepts in a *target* ontology. The resulting set of mappings is called an *alignment* [2]. As ontologies increase in number and size, automatic matching algorithms, which we call *matchers*, become not only important but absolutely necessary. However, in real-world scenarios, and even in the systematic ontology matching benchmarks of the Ontology Alignment Evaluation Initiative (OAEI), alignments are neither correct nor exhaustive when compared against a *gold standard*, also called *reference alignment*, created by domain experts.

By virtue of our collaboration with expert users in the geospatial domain [3], we came to realize the importance that they give to interacting with the matching process and to understanding why some mappings are present. In particular, users want to understand the provenance of the mappings, that is, they want to identify the reason why a mapping is part of an alignment. This is especially important when combining several matchers that evaluate different ontology features, a common technique in ontology matching.

We have therefore incorporated in the AgreementMaker system [4] a semi-automatic matching strategy in which domain experts are directly involved in an iterative feedback loop to validate mapping suggestions. At each iteration, the user is asked to validate one or more mappings. Our main objective is to improve the matching result as much as possible, while minimizing the user's effort. In order to do this, we answer the following two questions:
$Q1$: "Which specific candidate mappings should be presented to the user for validation?", and
$Q2$: "How can the user feedback be exploited to improve the existing alignment?"

We have developed visual analytic methods with the dual purpose of showing provenance of the mappings and of managing the overall matching process. To validate our approach, we conducted experiments using the OAEI benchmarks. The iterative feedback loop and the visual analytic methods described in this paper upgrade AgreementMaker by introducing new features that have not been presented elsewhere.

The importance of user feedback has been recently emphasized [5]. The "pay-as-you-go" approach [6] is based on the principle that integration at a large scale needs to be an ongoing process; it supports user feedback to improve upon initially found mappings over time. Two other approaches consider user feedback in ontology matching [7], [8]. However, in one of them [8] mappings that were either validated or refuted by the user are not remembered by the system. In fact, an infinite loop could potentially occur. In the other approach [7], there is no selection of candidate mappings with the aim of maximizing the information gain (e.g., by extrapolating to other mappings) in each iteration. As for our visualization component, it differs from that of a recent approach [9] as follows: (1) it displays the results of several matchers, not of a single matcher; (2) it is meant to be an integral part of the iterative matching process, not an *a posteriori* analytical tool.

II. Feedback Loop Core Principles

Given a source ontology S with m concepts, a target ontology T with n concepts, and l matchers, for each matcher $M_k, 1 \leq k \leq l$, we define an $m \times n$ similarity matrix Σ_k where the value $\sigma_k(i,j)$ of each element is in the interval $[0,1]$. We can then define a *signature vector* $\vec{v}_{i,j}$, where $\vec{v}_{i,j}(k) = \sigma_k(i,j), 1 \leq k \leq l$.

When determining the final alignment, a mapping cardinality and a similarity value threshold must be set, below which two concepts are not considered similar. Then an optimization algorithm is run to select the final alignment so as to maximize the overall similarity [10]. Similarly to several of the OAEI tracks, we adopt a 1-1 cardinality, meaning that each concept in the source (resp. target) ontology must be mapped to at most one single element in the target (resp. source) ontology. The final alignment is an $m \times n$ matrix A such that each row and each column of the matrix have a single non-zero element representing a *mapping* $(s_i, t_j, \sigma(s_i, t_j))$ between concept s_i in the source ontology and concept t_j in the target ontology, with similarity $\sigma(s_i, t_j)$ (a combination of $\sigma_k(i, j), 1 \leq k \leq l$ [10]) greater than the threshold.

Our feedback mechanism is based on the idea that the signature vector is key to the selection of which candidate mappings are presented next to the user as well as how that feedback is propagated to validate other mappings without presenting them to the user (so as to avoid presenting all possible mappings). We also do not present mappings that cannot be part of the final alignment due to cardinality constraints.

Taking into consideration question $Q1$ of Section I and the definition of the signature vector $\vec{v}_{i,j}$, we define a *disagreement* metric among the k similarity values. That is, if the matchers mostly agree on that mapping, then disagreement is low, but is otherwise high when the matchers produce different outcomes for that mapping. We select candidate mappings to be presented to the user at each step by ranking the mappings using the *Disagreement-based Top-k Mapping Selection* method.

As for question $Q2$, we use a *vector similarity* metric to *cluster* mappings based on their respective signature vectors. Once the user validates a mapping by confirming or refuting it, we *propagate* that information to those mappings whose signature vectors are most similar to the one associated with that mapping. This process uses the *Signature-based Mapping Clustering* and *Similarity Update and Propagation* methods.

Disagreement-based Top-k Mapping Selection. Different matchers consider different, possibly orthogonal, features of the ontologies to be aligned. Mappings over which matchers disagree upon are those mappings whose concepts are evaluated with high similarity according to some features and with low similarity according to other features. For this reason, these mappings represent potentially critical matching patterns. For example, concepts can be similar when evaluated by a syntactic matcher but dissimilar when evaluated by a structural matcher. User feedback can decide which criteria to favor when validating a mapping.

Given a signature vector $\vec{v}_{i,j}$, we define our *disagreement* metric as the variance of the similarity values in that vector. All mappings are then ranked according to this metric and the top-k mappings are selected for user feedback.

Signature-based Mapping Clustering. Given a user validated mapping and its corresponding signature vector, our approach identifies similar mappings by adopting a double threshold $th = (th^{\uparrow}, th^{\downarrow})$, with $th^{\uparrow}$ and $th^{\downarrow}$ being two positive rationals in the interval $[0, 1]$. For each dimension of the signature vector, we create a set of mappings whose similarities are within the threshold th of the similarity of the validated mapping. Having done this for all l-dimensions, we then take the intersection of the sets in order to select those mappings that belong to the cluster associated with the validated mapping.

Similarity Update and Propagation. Once a mapping is confirmed by the user it is added to the set of correct mappings; conversely, if it is refuted by the user, it is added to the set of incorrect mappings. To propagate the user's feedback on a mapping, we reward or penalize the similarity value associated with every mapping in the same cluster by a linear function that increases (resp. decreases) the similarity values; the similarity of mappings that have been validated (resp. refuted) is set to 1 (resp. 0) and not updated anymore.

III. System Architecture

As shown in Figure 1, our system consists of seven components, described in detail below.

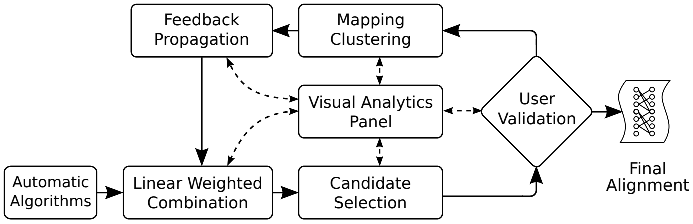


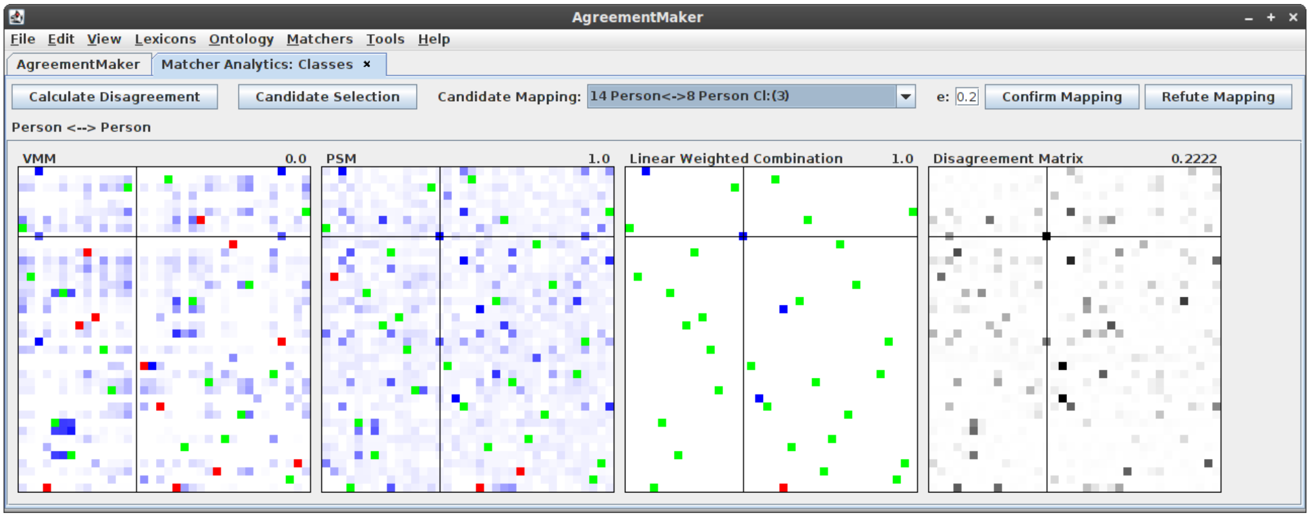
Fig. 1. System architecture.

Automatic Matching Algorithms. The automatic matchers include four syntactic-based matchers, among which are the *Advanced Similarity Matcher (ASM)*, the *Parametric String-based Matcher (PSM)*, and the *Vector-based Multi-word Matcher (VMM)* [4]. We also include a structural and a syntactic matcher, namely the *Iterative Instance and Structural Matcher* and the *Lexical Synonym Matcher* [11]. We run these automatic matchers in parallel and the result of each matcher k is stored in a similarity matrix Σ_k as defined in Section II.

Linear Weighted Combination (LWC). In this component, the similarity matrices of the automatic matchers are linearly combined [10] using weights determined by the *local confidence quality measure* [4]. The combined similarity value for each mapping is stored in the corresponding element of the similarity matrix Σ_{LWC}. An alignment is selected from this matrix that includes the best mappings. Similarity updates are performed on this matrix.

Candidate Selection. This component is responsible for the selection of the mappings that are presented to the user. The mappings that a user has validated in previous iterations are filtered out from the candidate mapping selection, so that the same mapping will not be validated twice. The *Disagreement-based Top-k Mapping Selection* mechanism is used to identify the k candidate mappings that are presented to the user. In addition, the user can add other mappings by inspection of the *Visual Analytics Panel* (Figure 2).

User Validation. This component allows the user to provide feedback on one or more candidate mappings. Candidate mappings are displayed in the *Visual Analytics Panel*, so as to facilitate the user's understanding of the corresponding signature vector. The system flags the mappings that have been confirmed or refuted and updates the alignment produced by the LWC component.


AgreementMaker
File Edit View Lexicons Ontology Matchers Tools Help
AgreementMaker
Matcher Analytics: Classes ×
Calculate Disagreement
Candidate Selection
Confirm Mapping
Refute Mapping
VMM 0.0
PSM 1.0
Linear Weighted Combination 1.0
Disagreement Matrix 0.2222

Candidate Mapping: 14 Person<->8 Person Cl:(3)

e: 0.2

Person <--> Person

Fig. 2. Visual Analytic Panel (partial view). The top toolbar controls the matching process. Colored squares represent mappings that are correct (green), missed (red), or falsely positive (blue). The intensity of the colors represents the similarity value in the range [0,1]. The overall panel highlights a vector of points for the same mapping. The two leftmost plots represent the similarity matrices for the VMM and PSM matchers. The third plot represents the Σ_{LWC} matrix. The rightmost plot represents the disagreement matrix among the matchers in shades of grey.

Mapping Clustering. This component associates a user-validated mapping with a set of similar mappings by adopting the *Signature-based Mapping Clustering* method. The boundary of the cluster can be adjusted by changing the clustering threshold th. Increasing the threshold will cluster increasingly dissimilar mappings, while decreasing the threshold will propagate the user feedback to fewer mappings, therefore a compromise must be found. The mappings belonging to the current cluster are shown in the *Visual Analytics Panel*; these mappings will be affected by the user's feedback, thus guiding the selection of an appropriate threshold.

Feedback Propagation. This component is responsible for the propagation of the user feedback on one mapping to other mappings using the *Similarity Update and Propagation* method.

Visual Analytics Panel. This panel, which is shown in Figure 2, is integrated in the user interface of AgreementMaker. Its visualization and control functionality assists users at each iteration of the feedback loop. A plot that represents the similarity matrix for each matcher is depicted, so as to give an overview of the distribution of the similarity values in the space of possible mappings. The reference alignment that is being progressively built is overlaid on each matcher's matrix plot for comparison; when users select a particular mapping, it is emphasized in every matrix plot. In this way, a comparative analysis of the corresponding signature vector is possible. Upon selection of a mapping, its cluster can also be visualized in each matrix plot. The analytics panel also controls the execution of the user feedback loop, with each step of the process invokable by clicking the appropriate button. Candidate mappings are displayed allowing the user to confirm or refute individual candidate mappings. For each validated candidate mapping, the user feedback is propagated to its cluster and candidate mappings are recomputed. At any step of the process, users can explore alignments using the side by side ontology view of the user interface (see Figure 3).

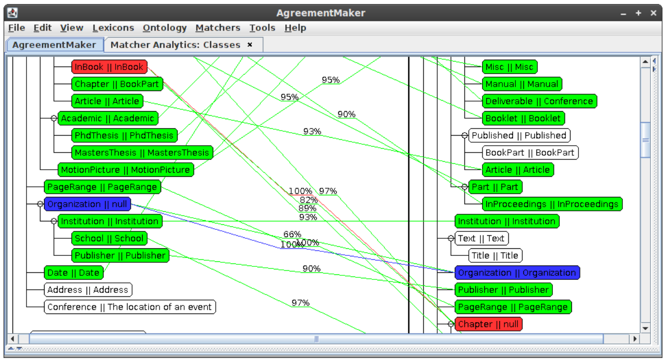


Fig. 3. AgreementMaker user interface: ontology view.

IV. Experimental Evaluation

Our experiments are targeted to evaluate the effectiveness of the core principles of our approach. In particular, we show that the selection of candidate mappings presented to the user and the propagation of the user's feedback, when paired together, result in a significantly improved final alignment. In our experiments, the user validation is simulated at every iteration via the reference alignment. We used the OAEI benchmarks ontology sets, which consist of real-world bibliographic reference ontologies that include BibTeX/MIT, BibTeX/UMBC, and INRIA, and their reference alignments.

Figure 4 shows that the top 60 (1.1%) of the most disagreed upon mappings make up 60% of the reference alignment, proving that there is a strong correlation between mappings that are disagreed upon and those that are relevant for presentation to the user for feedback. Figure 5 shows an average F-Measure gain of 7.2% as a result of the similarity propagation method. This is a sizable gain considering that we started from an

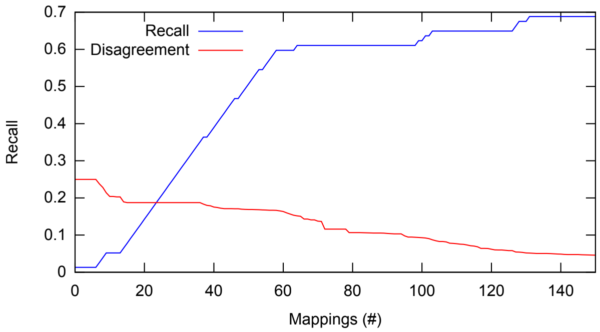


Fig. 4. Candidate selection evaluation.

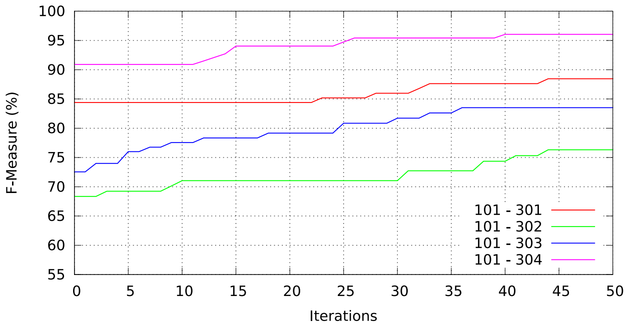


Fig. 5. Similarity propagation evaluation on the OAEI test cases.

already high average F-Measure of 80.6%, which was obtained using the automatic matchers, and that it was realized with 50 iterations, which represent only 1.26% of the mapping search space.

The *Visual Analytics Panel* of Figure 2 brings a whole new light to the matching process allowing users to discover matching patterns that were previously hidden in the complexity of the process. In particular, this panel greatly helped us in pinpointing correct and missed mappings in the course of our own experiments.

V. Demonstration Scenario

We demonstrate our ontology matching system by having users match OAEI and Linked Open Data ontologies (e.g., DBpedia) [12]. The users can follow step by step the flow of the interactive feedback-based matching method (see Figure 1), while visualizing and controlling each step from the user interface of AgreementMaker.

A first set of mappings is automatically computed by the automatic matchers. As the first step to completing or correcting this set of mappings, a list of top-k most disagreed-upon mappings is presented. Users can visualize a candidate mapping's concepts in the ontology view, and also visualize the mapping signature of each candidate mapping in the analytics panel in order to make a decision whether to confirm or refute the candidate mapping. In addition, users can also visualize the mapping cluster to which their feedback is propagated and compare it with the reference alignment they are building. They can decide whether to propagate their feedback, to adjust the cluster size via the threshold, or to validate another mapping. Once a choice is made, the effects of the user-provided feedback are highlighted in the analytics panel. To informally demonstrate the importance of the visual analytics panel, during the demonstration users will be invited to use the system with and without that panel and draw their own conclusions. Another demo mode will simulate the next step based on the reference alignment. At each step the chosen mapping will be highlighted and full interaction with the user interface will also be allowed.

Acknowledgments

The work on this paper was partially supported by NSF Awards IIS-1143926 and IIS-0812258 and by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL) contract number FA8650-10-C-7061. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government. Matteo Palmonari's work was performed in part while visiting the ADVIS Lab.

References

[1] E. Rahm and P. A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," *VLDB Journal*, vol. 10, no. 4, pp. 334–350, 2001.

[2] J. Euzenat and P. Shvaiko, *Ontology Matching*. Heidelberg (DE): Springer-Verlag, 2007.

[3] I. F. Cruz and W. Sunna, "Structural Alignment Methods with Applications to Geospatial Ontologies," *Transactions in GIS, Special Issue on Semantic Similarity Measurement and Geospatial Applications*, vol. 12, no. 6, pp. 683–711, December 2008.

[4] I. F. Cruz, F. Palandri Antonelli, and C. Stroe, "AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies," *PVLDB*, vol. 2, no. 2, pp. 1586–1589, 2009.

[5] K. Belhajjame, N. W. Paton, A. A. A. Fernandes, C. Hedeler, and S. M. Embury, "User Feedback as a First Class Citizen in Information Integration Systems," in *CIDR Conference on Innovative Data Systems Research*, 2011, pp. 175–183.

[6] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go User Feedback for Dataspace Systems," in *ACM SIGMOD International Conference on Management of Data*, 2008, pp. 847–860.

[7] S. Duan, A. Fokoue, and K. Srinivas, "One Size Does Not Fit All: Customizing Ontology Alignment Using User Feedback," in *International Semantic Web Conference (ISWC)*, ser. Lecture Notes in Computer Science, vol. 6496. Springer, 2010, pp. 177–192.

[8] F. Shi, J. Li, J. Tang, G. Xie, and H. Li, "Actively Learning Ontology Matching via User Interaction," in *International Semantic Web Conference (ISWC)*, ser. Lecture Notes in Computer Science, vol. 5823. Springer, 2009, pp. 585–600.

[9] E. Peukert, J. Eberius, and E. Rahm, "AMC—A Framework for Modelling and Comparing Matching Systems as Matching Processes," in *IEEE International Conference on Data Engineering (ICDE)*, 2011, pp. 1304–1307.

[10] I. F. Cruz, F. Palandri Antonelli, and C. Stroe, "Efficient Selection of Mappings and Automatic Quality-driven Combination of Matching Methods," in *ISWC International Workshop on Ontology Matching (OM)*, ser. CEUR Workshop Proceedings, vol. 551, 2009, pp. 49–60.

[11] I. F. Cruz, C. Stroe, M. Caci, F. Caimi, M. Palmonari, F. Palandri Antonelli, and U. C. Keles, "Using AgreementMaker to Align Ontologies for OAEI 2010," in *ISWC International Workshop on Ontology Matching (OM)*, ser. CEUR Workshop Proceedings, vol. 689, 2010, pp. 118–125.

[12] I. F. Cruz, M. Palmonari, F. Caimi, and C. Stroe, "Towards "On the Go" Matching of Linked Open Data Ontologies," in *IJCAI Workshop Discovering Meaning On the Go in Large & Heterogeneous Data (LHD)*, 2011, pp. 37–42.