IBM Research

A Clustering-based Approach to Ontology Alignment

Songyun Duan, Achille Fokoue, Kavitha Srinivas: IBM Research

Brian Byrne: IBM Software Group

Outline

- Motivation
- Overview of the clustering-based approach
- Measures to test our central hypothesis
- Experimental evaluation
- Conclusion

Motivation

- Numerous (semi-)automated alignment systems developed in recent years
  - e.g., Lily, ASMOV , Anchor-Flood, RiMOM [13].
  - Combination a large set of similarity functions on lexical, semantic and structural features to align ontologies
  - Approach important and effective for many cases of ontology alignments
- What if none of the similarity functions adequately capture the nature of the alignment?
- Not just a theoretical hypothetical:
  - Frequent situation encounters by IBM consultants
  - Mapping IFW (Information FrameWork) vs CBM (Component Business Model)
    - Asset at the IT level vs Same Asset at the Business level
    - Different vocabularies: IT terms vs Business terms
    - Different structures: IFW hierarchical model vs CBM flat model
- Traditional automated ontology alignment fail abysmally (precision ~1%).

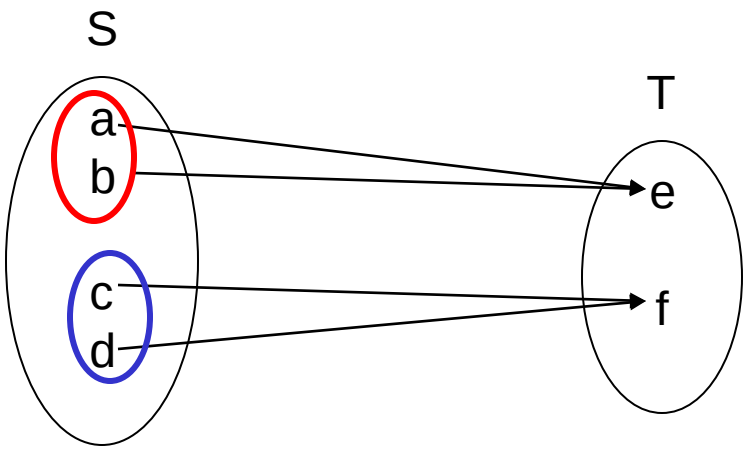
Motivation (II)

- Only alternative: Rely on a domain expert for alignment
- Our research question: how can these high quality manual mappings be re-used for new alignments?
  - E.g. when the two models evolve
  - E.g. when the same model needs to be align with new models.
- Focus on many-one (one-many) alignments

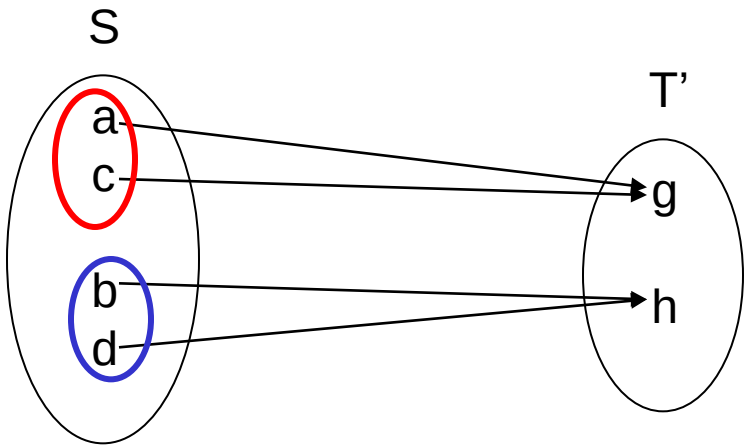
Outline

- Motivation
- Overview of the clustering-based approach
- Measures to test our central hypothesis
- Experimental evaluation
- Conclusion

Overview of Clustering-based Ontology Alignment (Stability Hypothesis)

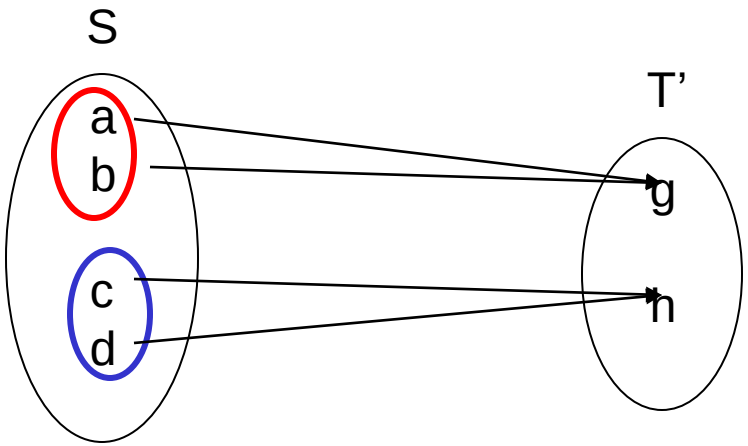


P_s ={ {a,b}, {c,d}}



P_s ={ {a,c}, {b,d}}

Stability Hypothesis: ***The partitions of a source ontology (based on alignment results with different target ontologies) are stable across ontology alignments***



Overview of Clustering-based Ontology Alignment (Leveraging previous alignment results)

- Assuming the Stability Hypothesis holds
- Simple mechanism to leverage existing alignment S → T for S → T'
  - Generate a partition (i.e., a set of clusters) of S, Ps, from the alignment of S → T
  - To perform the alignment task of S → T',
    - Individual entities in S are NOT independently with the entities in T',
    - More efficient and more accurate to match a cluster of entities in Ps to the entities in T'.
  - The intuition: entities in one cluster are expected to match to the same entity in T'
  - Particularly efficient for ontology evolution scenario

Overview of Clustering-based Ontology Alignment (Real Examples)

IFW	CBM
Provide FMO Transaction Reconciliation	Account Reconciliation
Request Amended Counterparty Confirmation	Account Reconciliation
Accumulate Futures Transaction Values	Account Reconciliation
Analyze FMO Transaction Details	Account Reconciliation
Compare FMO Transaction Details	Account Reconciliation
Verify FMO Transaction Details	Account Reconciliation

Table 1. Example of an IFW cluster based on manual alignment to CBM

- Most entities in the IFW cluster show little to no lexical or structural similarity between themselves or with the target CBM Entity
- Standard approach to map IFW to CBM produce extremely poor results
- Semantic similarity between IFW entities in the cluster was indirectly identified by the domain experts

Overview of Clustering-based Ontology Alignment (Real Examples II)

Mouse Anatomy	Brenda Tissue
intestine (no synonym)	intestine (synonyms: bowel, gut)
bowel (no synonym)	intestine (synonyms: bowel, gut)
gut (no synonym)	intestine (synonyms: bowel, gut)

Table 2. Example of a Mouse Anatomy cluster based on lexical alignment to Brenda Tissue Ontology

- No lexical similarity between entities in MA cluster
- However, lexical similarity between MA entities & BT entity
- BTO serves as a dictionary look up that uncovers the semantic similarity between *intestine*, *bowel*, and *gut*.
- Use this uncovered semantic similarity in the next MA alignment.

Outline

- Motivation
- Overview of the clustering-based approach
- Measures to test our central hypothesis
- Experimental evaluation
- Conclusion

Testing the Stability Hypothesis (Partition similarity measures)

- Evaluate the similarity between the partitions of the same ontology
- Two alignments: $S \rightarrow T_1$ and $S \rightarrow T_2$.
  - Two partitions of S: $P_{s,1} = \{C_1, C_2, \ldots, C_m\}$ and $P_{s,2} = \{C_1', C_2', \ldots, C_n'\}$
- Challenge: define an appropriate measure to evaluate the similarity of $P_{s,1}$ and $P_{s,2}$
- Two similarity measures similar to similarity metrics for strings:
  - Measure I: Jaccard Similarity on Entity Pairs
  - Measure II: Partition Edit Distance
- Direct measure to evaluate the actual quality of mappings generated based on the clustering information:
  - Measure III: Mapping Quality

Measure I: Jaccard Similarity on Entity Pairs

- $S \rightarrow T_1$: $P_{S,1}$ ={{a, b}, {c, d, e}}
  - $P_{S,1}$' = {{a,b}, {c,d}, {c,e}, {d,e}} (P_S' <-> P_S)
- $S \rightarrow T_2$: $P_{S,2}$ ={{a, b,c}, {d, e}}
  - $P_{S,2}$' = {{a,b}, {a,c}, {b,c}, {d,e}}
- $PSim_I$ ($P_{S,1}$, $P_{S,2}$) = Jaccard($P_{S,1}$' , $P_{S,2}$')

  = |$P_{S,1}$' ∩ $P_{S,2}$'| / |$P_{S,1}$' ∪ $P_{S,2}$'|

  = |{{a,b}, {d,e}}| / | {{a,b}, {c,d}, {c,e}, {d,e},{a,c}, {b,c}} | =2/6=0.33
- $PSim_I$ has the desired properties:
  - Symmetric
  - [0,1] range
- $PSim_I$ captures the effect of big clusters in a partition
  - big clusters generate entity pairs that are exponential in size to cluster size

Measure II: Partition Edit Distance

- $S \rightarrow T_1$: $P_{S,1}$ ={{a, b}, {c, d, e}}
- $S \rightarrow T_2$: $P_{S,2}$ ={{a, b,c}, {d, e}}
- How can we transform $P_{S,1}$ into $P_{S,2}$ using two types of operations: Split and Merge?
  - Split of a cluster C => two disjoint sets C_1 and C_2 s.t $C = C_1 \cup C_2$
  - Merge of C_1 , C_2 => $C_1 \cup C_2$
- $P_{S,1}$ -> $P_{S,2}$:
  - a Split operation on the cluster *{c, d, e}* => *{c}* and *{d, e}*
  - a Merge operation *{c}* and *{a, b}* => *{a,b,c}*
- Definition: *The edit distance between two partitions P_1 and P_2, ED(P_1, P_2), is the length of the shortest edit path composed of Splits and Merges from P_1* to *P_2. E.g. ED($P_{S,1}$, $P_{S,2}$) = 2 . ED is symmetric*
- $PSim_I$ ($P_{S,1}$, $P_{S,2}$) = 1 - *ED($P_{S,1}$, $P_{S,2}$)/|S| (=1-2/5 = 0.6)*

Measure III: Mapping Quality

- Direct measure of the actual quality of mappings which generated based on the clustering information
- Cluster-based mapping process
  - Generate a partition P_1 of the source ontology S based on the mapping result from S to a target ontology T_1
  - For a new alignment task from S to another target ontology T_2, generate the mappings as follows:
    - For each cluster C in the partition P_1, randomly pick one entity s from C and find the mapped entity t in T_2;
    - Generalize the mapping to other entities in the same cluster, with the mappings being $\{\langle s'; t\rangle | s' \in C\}$.
- Standard precision and recall measures of this mapping process
  - Precison = $|M \cap M_{GS}| / |M|$
  - Recall = $|M \cap M_{GS}| / |M_{GS}|$
- Efficiency: the human effort estimated as the number of mappings that require human input
  - Efficiency = 1 – (|NonSingletonClusters| / |{e| e in a non singleton cluster}|)

Discussion on the previous measures

- Measure I: Jaccard Similarity on Entity Pairs
  - Very sensitive to difference in partition granularity
    - P_1 = {{a, b, c, d}} and P_2 ={{a,b}, {c,d}}
      - P_2 obtained from a target T_2 at a finer granularity
      - Small Jaccard similarity (1/3 =0.33)
- Measure II: Partition Edit Distance
  - Insensitive to difference in partition granularity
    - ED(P_1, P_2) =1
    - Edit similarity = 1-1/4 = 0.75
- Measure III: Mapping quality
  - Reliability of the partition information for end use.

Outline

- Motivation
- Overview of the clustering-based approach
- Measures to test our central hypothesis
- Experimental evaluation
- Conclusion

Experimental Evaluation of the Partitioning Stability: IFW - CBM: Ontology Evolution Scenario

- Two high-quality manual alignments: IFW – CBM_{v1} and IFW – CBM_{v2}

	CBM_{v1}	CBM_{v2}	CBM_{v1} ≡ CBM_{v2}	IFW
entities	65	120	37	2165

	IFW (CBM_{v1})	IFW (CBM_{v2})
clusters	62	111
avg size	34.92	19.5
total avg	25	

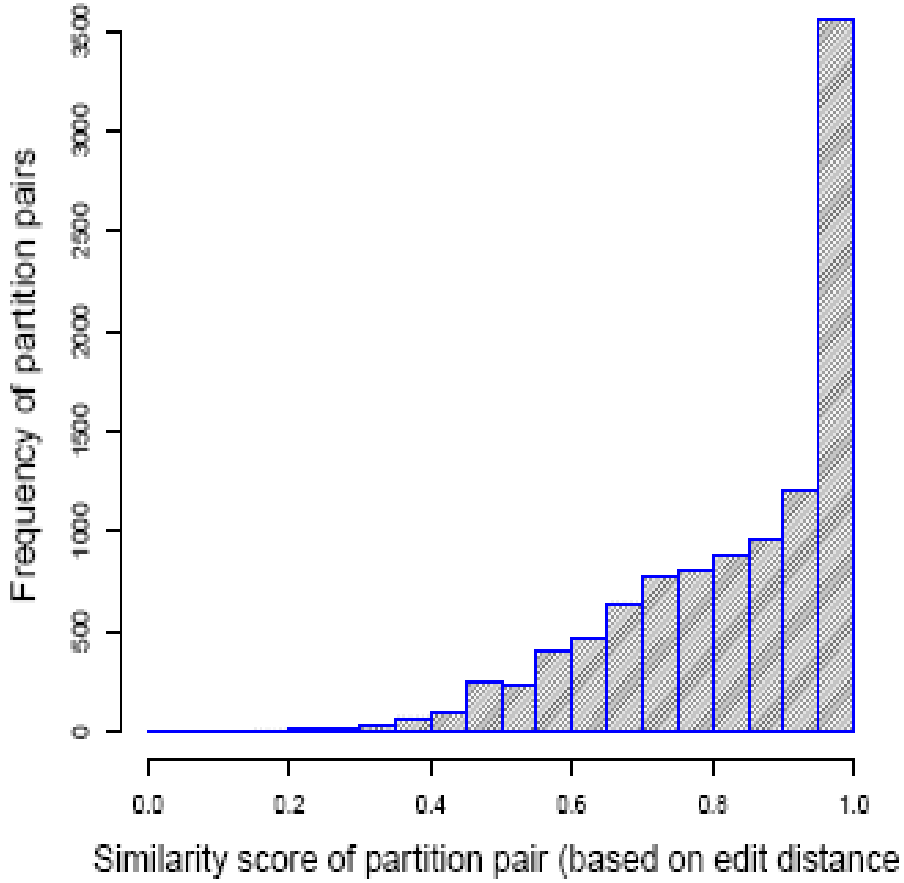
	Edit dist	Jaccard
IFW (CBM_{v1})- IFW (CBM_{v2})	0.89	0.53

	efficiency	precision
IFW vs CBM	0.95	0.78

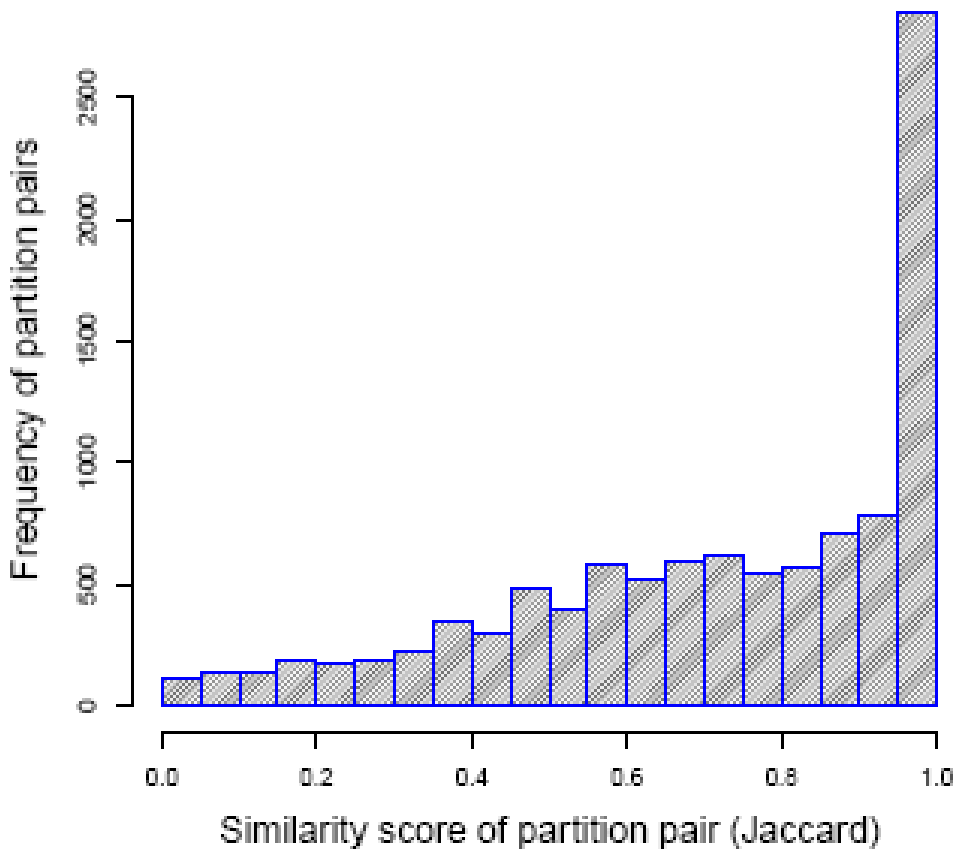
Experimental Evaluation of the Partitioning Stability: Large Scale Evaluation on BioPortal Ontologies

- BioPortal
  - 149 ontologies in life sciences domain
  - 9.3K ontology comparisons
  - 1.75 million matchings of elements (mostly lexically generated)
- For each ontology S mapped to k target ontologies
  - k partitions of S
  - Evaluate edit and jaccard similarity on all pairs of partitions
- Result: 10.4K pairs of partitions evaluated against the two similarity metrics
  - Average cluster size 2

Experimental Evaluation of the Partitioning Stability: Large Scale Evaluation on BioPortal Ontologies II



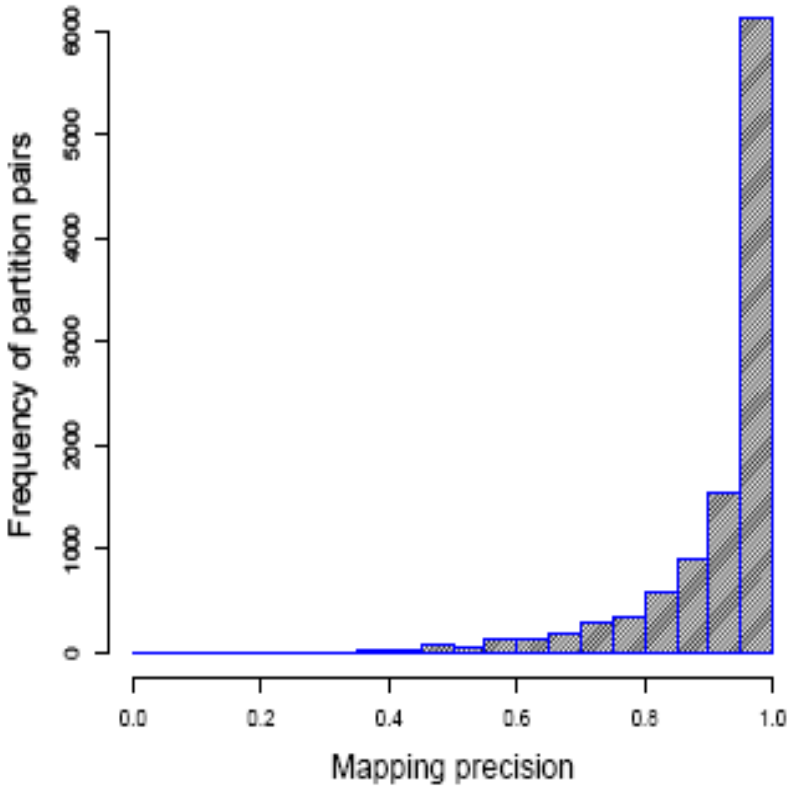
Mean = 0.84, Stdv = 0.12



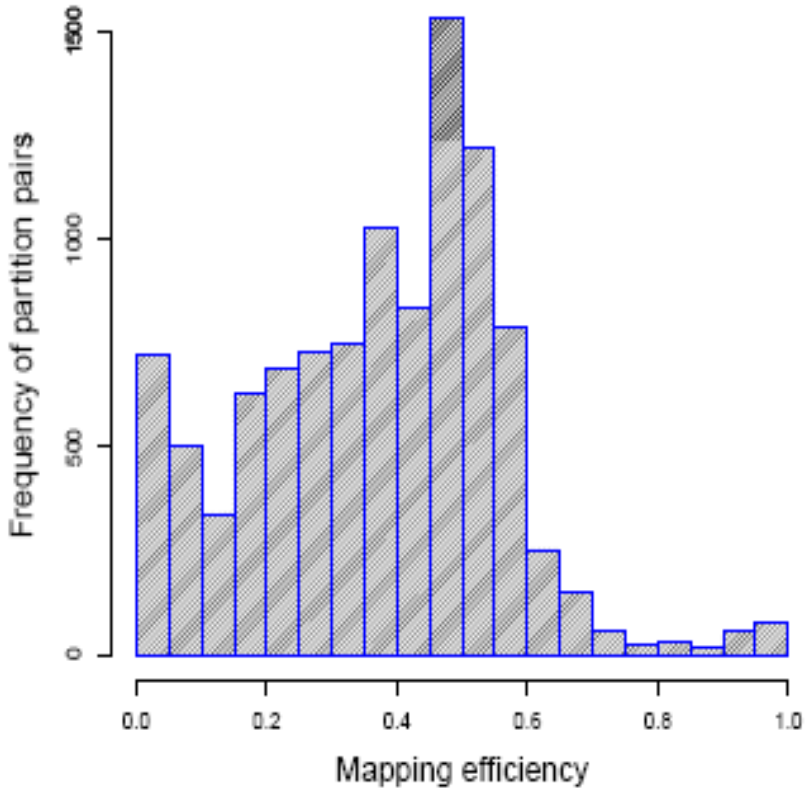
Mean = 0.72 Stdv = 0.26

Bottom line: Partitions of an ontology S are reasonably stable

Experimental Evaluation of the Mapping Quality: Large Scale Evaluation on BioPortal Ontologies II



Mean = 0.92, Stdv = 0.11



Mean = 0.37 Stdv = 0.19

Bottom line 1: Mapping precision confirms the viability of using clustering info

Bottom line 2: Modest efficiency due to small cluster size (avg 2)

Conclusion

- We present a novel technique to uncover, from existing many-to-one (or conversely, one-to-many) alignments, internal structures of related entities (i.e., clusters of entities) in ontologies.
- We show the stability of those clusters across alignments in two different domains (finance and healthcare & life sciences) and on both manually created mappings and automatically generated high-quality mappings.
- We describe how clusters discovered in existing many-to-one and one-to-many alignments can be exploited for performing new alignments, and evaluate the impact on both mapping quality (precision/recall) and mapping efficiency (saving in human effort).

THANKS!

Questions?

Contact: Achille Fokoue
Email: achille@us.ibm.com