sample evaluation of ontology-matching systems

Willem Robert van Hage
Antoine Isaac
Zarko Aleksovski

Vrije Universiteit Amsterdam

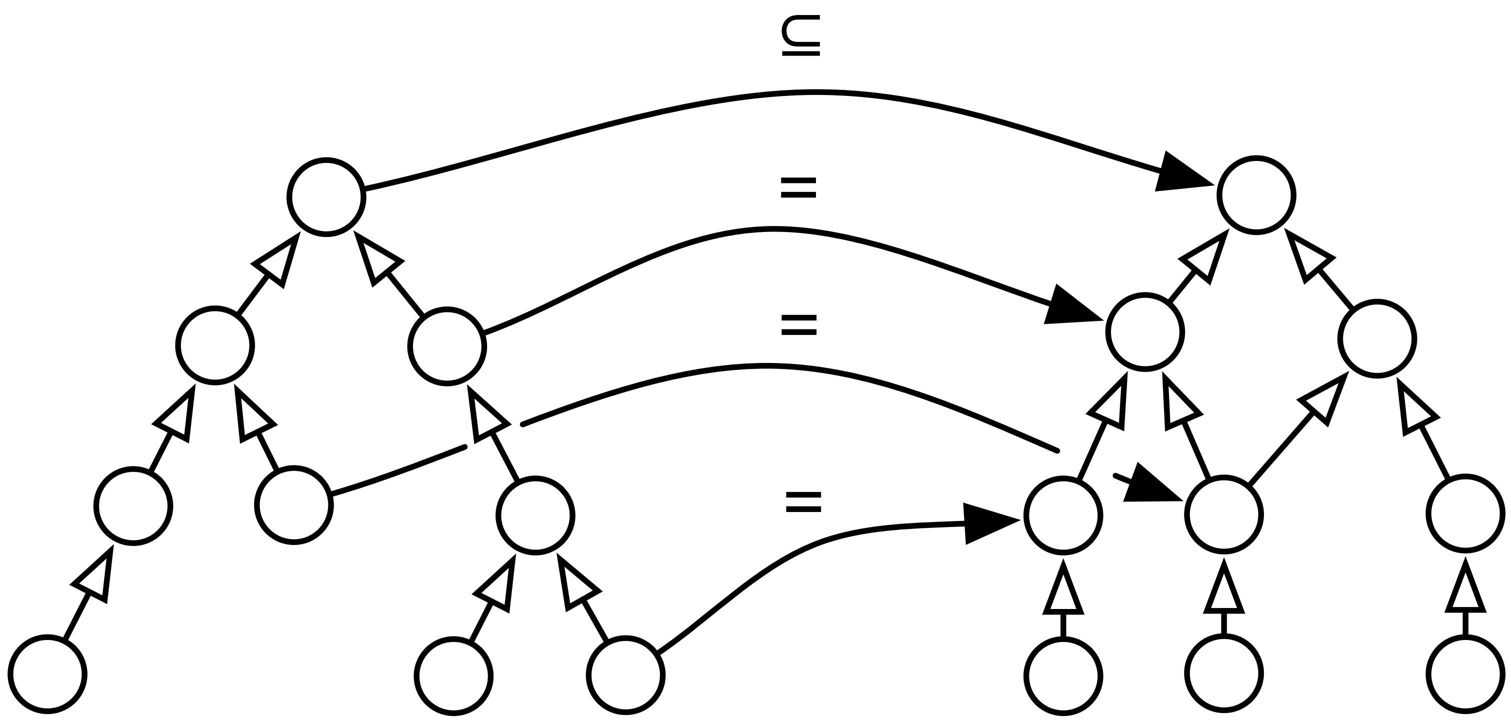
EON 2007 - Busan, Korea

overview

- situation: Ontology matching
- problem: Evaluation lacks link to application
- solution: Application-based evaluation
  - approach 1: End-to-end evaluation
  - approach 2: Alignment sampling

situation

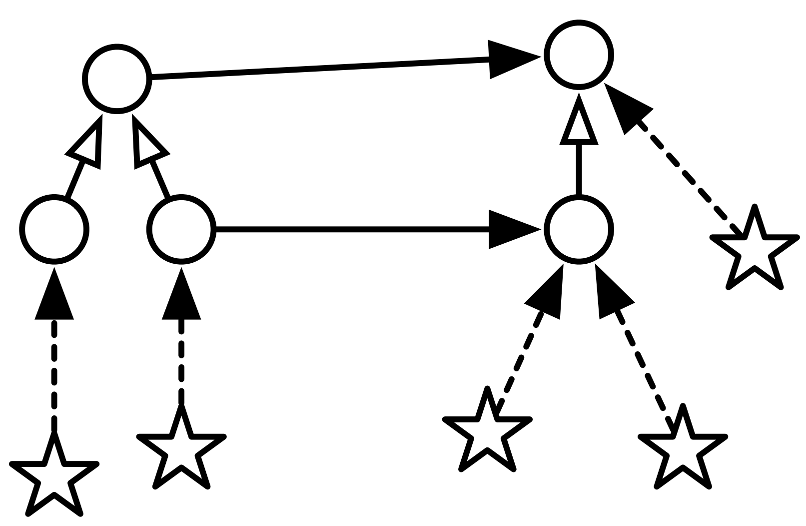
ontology matching



- Euzenat & Shvaiko:
  - Ontology matching produces a set of correspondences that is called an alignment
  - Mappings are one kind of correspondences

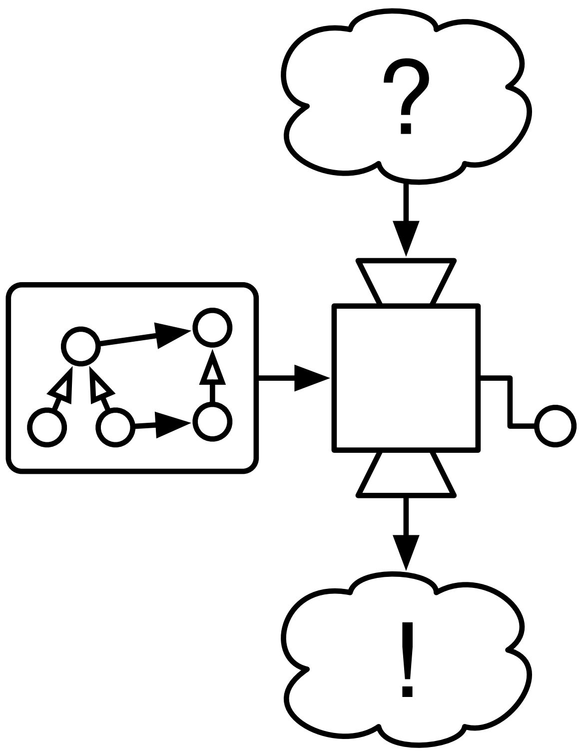
ontology matching

- Usually, the alignment is used to improve the performance of some information system
    - add more concepts
    - add more instances



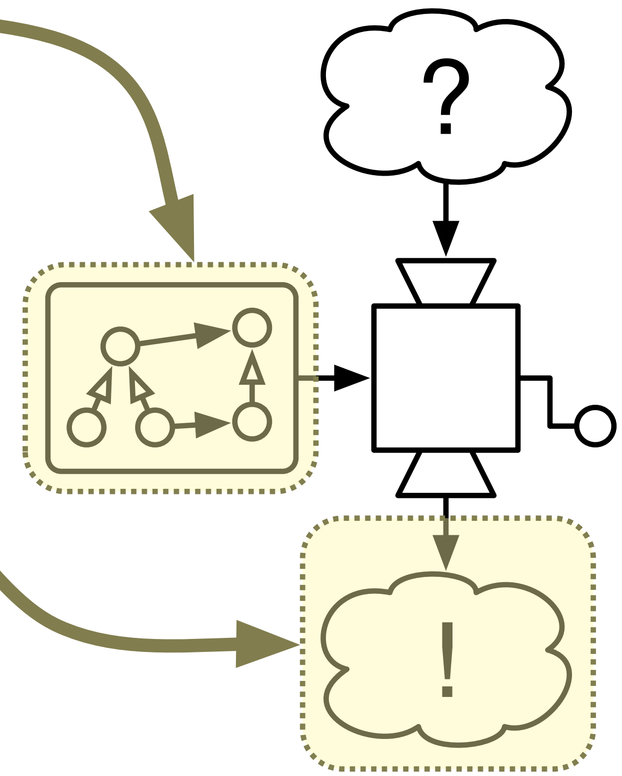
evaluation approaches

1. Assess correspondences
    - metric: e.g. percentage of objectively correct correspondences in the alignment
2. Assess system performance
    - metric: e.g. percentage of queries for which the alignment improves the quality of the application



evaluation approaches

1. Assess correspondences
   - metric: e.g. percentage of objectively correct correspondences in the alignment
2. Assess system performance
   - metric: e.g. percentage of queries for which the alignment improves the quality of the application



evaluation status quo

- Count correct correspondences to estimate Precision and Recall
- Usually, only one average number is provided per alignment
- Application of the alignment is largely ignored

problem

problems with status quo

- No evaluation of the benefits for users
- Only the correctness of the alignment is tested,
  not the relevance of various parts for the application
  - e.g. If you just need 10% of the alignment, but the matcher only finds the other 90% then 90% Precision is low.
- Average numbers smooth out important details
  - You want separate numbers for every matching relation and for correspondences in different domains
    e.g. geography, law, mechanics, taxonomy, etc.

problems with status quo

That equates to:

- Lack of time
- Lack of statistical foundation

solution

our propositions

- Propose easy end-to-end evaluation method to measure user statisfaction
  (approach 1)
- Provide statistical foundation for sampling that allows quicker evaluation and thus allows for more tests per case
  (approach 2)
- Provide statistical foundation for drawing conclusions based on various samples
  (approach 2)

approach 1

end-to-end evaluation

- Capture user satisfaction by formulating a number of queries that:
  - Represent every topic of interest
  - Fairly represent the commonness or rarity of each topic in actual usage
  - Fairly represent difficult and easy topics
- Pick a measure for user satisfaction based on the results (of the information system, i.e. the set/list of objects that is found)

stand-alone

end-to-end evaluation

- Real-life queries immediately reveal if a matching system can find the correspondences that solve the problem
- Variance of the results depends on the measure for user satisfaction that is used and the number of queries
- Analysis of the variance requires repetition of the experiment (expensive)
- Future work...

comparative

end-to-end evaluation

- We measure:
  the number of queries for which one information system (c.q. alignment) outperforms the other

- A system is significantly better if the number of improved queries is larger than can be expected “by chance”
- We reduce it to:
  determining **if** a coin is biased by flipping it *n* times heads: *X* better, tails: *Y* better we expect 50-50

comparative

end-to-end evaluation

- We use the Sign test
- S_+ is the total number of times X is better than Y
- X is significantly better if:

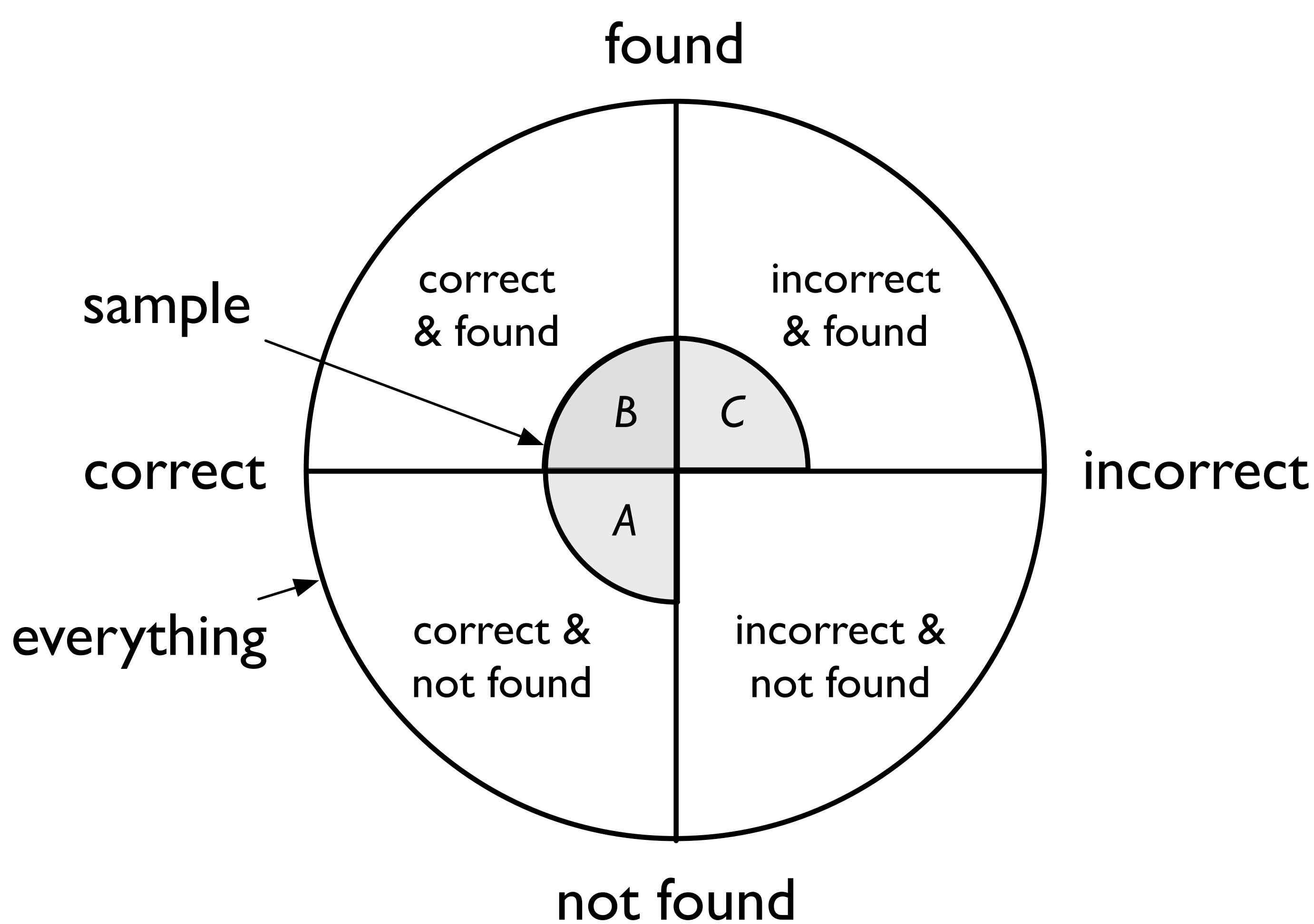
$$\frac{2 \cdot S_+ - n}{\sqrt{n}} > 1.96$$

query #	X better	Y better
1	✓	
2	✓	
3		✓
4	✓	
5		✓
...		
n		✓

approach 2

alignment sampling

- Measure Recall with sample $A \cup B$
- Measure Precision with sample $B \cup C$



approach 2

alignment sampling

- Construct Recall sample by making a set of true correspondences.
  - It does not matter how you derive these, as long as they are randomly selected from the set of Correct correspondences
  - In most cases creating an alignment between arbitrarily selected portions of the ontologies sufficiently approximates a random selection
- Construct Precision sample by making a set of found correspondences (i.e. run a matcher)
  - Take a random sample from the found correspondences

approach 2

alignment sampling

- We measure:
  the proportion of correct correspondences in a large set of correspondences
  (either for Precision or Recall, it doesn't matter)
- The proportion in the sample is an estimation of the true proportion.
  The error depends on the sample size and the actual true proportion, which we will never know exactly.
- We reduce it to estimating the bias of a coin by flipping it *n* times

stand-alone

alignment sampling

- For both Precision and Recall samples goes
  - The margin-of-error based on a sample of size *n* at a confidence level of 95% is at most (and usually less than)

$$\frac{1}{\sqrt{n}}$$

  - e.g. If 75% of a sample of correspondences of size 100 is correct then the margin-of-error is 0.1 = 10%
    i.e. the true value lies between 65% and 85% with a confidence of 95%
  - A sample of size 1000 gives a margin-of-error of 0.03 = 3%
    i.e. the true value lies between 72% and 78%

comparative

alignment sampling

- For both Precision and Recall samples goes
  - If system *X* is better than system *Y* with a confidence of 95% if the proportion of correct mappings differs by at least

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\frac{\hat{P}_A(1-\hat{P}_A)}{n} + \frac{\hat{P}_B(1-\hat{P}_B)}{n}}$$

  - e.g. With sample of correspondences of size 100 we can distinguish differences of at least 0.14 = 14% (like 70% and 84%) with a confidence of 95%
  - e.g. With a sample of correspondences of size 1000 we can distinguish differences of at least 0.04 = 4% (like 70% and 74%)

comparative

alignment sampling

- For both Precision and Recall samples goes
  - If system *X* is better than system *Y* with a confidence of 95% if the proportion of correct mappings differs by at least (upper bound at p = 0.5)

$$\frac{2}{\sqrt{2 \cdot n}}$$

  - e.g. With sample of correspondences of size 100 we can distinguish differences of at least 0.14 = 14% (like 70% and 84%) with a confidence of 95%
  - e.g. With a sample of correspondences of size 1000 we can distinguish differences of at least 0.04 = 4% (like 70% and 74%)

approach 2

alignment sampling

- For the non-simplified formula's of the variance and margin-of-error see the paper:

  `http://www.few.vu.nl/~wrvhage/papers/eon2007wrvh.pdf`

stratified

alignment sampling

- Partition the entire alignment into sets of correspondences called strata
- Each stratum should be a set of similar correspondences
  i.e. different matching relations or different topics that represent different usage
- Perform alignment sample evaluation for each stratum
- Combine the results to get the overall score by taking a weighted average

stratified

alignment sampling

- Benefits over plain alignment sampling:
  - Different performance measurements for parts of the alignment with a different purpose
  - The same total sample size gives a smaller margin-of-error
    (it removes the possibility that the differences accounted for by the stratification are accidentally ignored in the random sample)

stratified

alignment sampling

- For details see the paper:

  `http://www.few.vu.nl/~wrvhage/papers/eon2007wrvh.pdf` questions?